A Very Short Introduction

牛津通识读本

数学

Mathematics

[英国] 蒂莫西·高尔斯 / 著
刘熙 / 译

译林出版社

VISIT...

LANZAROTE Caliente.COM

A Very Short Introduction

牛津通识读本

数学

Mathematics

[英国] 蒂莫西·高尔斯 / 著
刘熙 / 译

译林出版社

图书在版编目（CIP）数据

数学/（英）高尔斯（Gowers，T.）著；刘熙译.——南京：译林出版社，2014.3

（牛津通识读本）

书名原文：Mathematics：A Very Short Introduction

ISBN 978-7-5447-4523-9

I.① 数… II.① 高… ② 刘… III.① 数学——研究 IV.① 01——0

中国版本图书馆CIP数据核字（2013）第242823号

Copyright © Timothy Gowers 2002

Mathematics was originally published in English in 2002.

This Bilingual Edition is published by arrangement with Oxford University Press and is for sale in the People's Republic of China only，excluding Hong Kong SAR，Macau SAR and Taiwan，and may not be bought for export therefrom.

Chinese and English edition copyright © 2013by Yilin Press，Ltd

著作权合同登记号图字：10-2007-046号

书　名 数学

作　者 ［英国］ 蒂莫西·高尔斯

译　者 刘　熙

责任编辑 何本国

原文出版 Oxford University Press，2007

出版发行 凤凰出版传媒股份有限公司 译林出版社

集团地址 南京市湖南路1号A楼，邮编：210009

集团网址 http://www.ppm.cn

出版社地址 南京市湖南路1号A楼，邮编：210009

电子信箱 yilin@yilin.com

出版社网址 http://www.yilin.com

版　次 2014年3月第1版 2014年3月第1次印刷

书　号 ISBN 978-7-5447-4523-9

目录

前言
序言
第一章 模型
第二章 数与抽象
第三章 证明
第四章 极限与无穷
第五章 维度
第六章 几何
第七章 估计与近似
第八章 常见问题
索引

前言

20世纪初，伟大的数学家大卫·希尔伯特发现，有很多数学中的重要论点在结构上十分类似。他意识到，在适当的广义范畴下，这些论点事实上可以视为等同。与此类似的一系列发现为一个崭新的数学分支开启了大门。而这一新领域中的一个核心概念——希尔伯特空间——正是以希尔伯特的名字来命名的，这一概念使许许多多的现代数学研究变得清晰，范围之广包括了从数论直到量子力学各个分支，以至于如果你对希尔伯特空间的基本理论一无所知，你就根本不能算是一名受过良好教育的数学家。

那么，什么是希尔伯特空间呢？在典型的高校数学课程中，它被定义为“完备的内积空间”。修读这样一门课程的学生，理应从先修课程中了解到，所谓“内积空间”是指配备了内积的向量空间，而所谓“完备”是指空间中任意柯西列都收敛。当然，要想理解这样的定义，学生还必须知道“向量空间”、“内积”、“柯西列”和“收敛”的定义。就拿其中一个举例来说（这还并不是最长的一个）：序列x_1，x_2，x_3，…若满足对于任意正数ε，总存在整数N，使得对于任意大于N的整数p和q，x_p与x_q间的距离不大于ε，则称这个序列为柯西列。

简言之，如果你希望了解希尔伯特空间是什么，你就必须首先学习并且消化一系列由低到高、等级分明的较低级概念。毫无疑问这需要耗费时间和精力。对于许多最重要的数学思想来说都是这样。有鉴于此，要写一本书提供对数学的简单易懂的介绍，其所能达到的目标就极为有限，更何况这本书还需要写得很短。

我没有选择用更聪明的办法绕着这个难题走，而是集中关注数学交流中另一重完全不同的障碍。这重障碍并非技术性的，而更多属于哲学性质的。它区分开了两种人：一种人乐于接受诸如无穷大、负一的平方根、第二十六维和弯曲空间这样的概念，另一类人则觉得这些概念荒诞不经。其实无须沉浸在技术细节中，依然有可能坦然接受这些思想，我将努力表明如何做到这一点。

如果说这本书要向你传达什么信息的话，那就是——我们应当学习抽象地思考，因为通过抽象地思考，许多哲学上的困难就能轻易地消除。在第二章里，我将详细说明什么是抽象的方法。第一章中则考虑我们更熟悉、与日常更相关的抽象：从现实世界的问题中提取核心特征，从而将其转化为数学问题的过程。第三章中我将讨论什么叫作“严格的证明”。这前三章是关于一般性的数学的。

之后我将讨论一些更加具体的课题。最后一章与其说是关于数学的，不如说是关于数学家的，因此会跟前几章有些不同。我建议你在读过第二章后再阅读后续章节。除此以外，这本书已经尽量做到不受先后顺序影响——在任何章节中，我并没有假定读者已经理解并记住了先前的内容。

读这本书并不需要太多的预备知识，英国GCSE课程[1]或同等水平即可。不过我假定读者具有一些兴趣，而不是需要靠我去大力宣扬。因此，我在书中没有用到趣闻轶事、漫画、惊叹号、搞笑的章节标题或者曼德布罗特集合[2]的图片。我同样避免了混沌理论、哥德尔定理等内容：与它们在当前数学研究中的实际影响相比，这些内容在公众的想象

中所占的比例已经过大，而且其他图书已经充分地阐释了这些内容。我所选择的内容都是很普通的，详细地去讨论，以说明怎样通过一种更深刻的方式来理解它们。换言之，我的目标在深不在广，在于向读者传达主流数学的魅力，让读者体会到它的不言而喻。

感谢克雷数学研究所和普林斯顿大学在我写作此书期间对我的支持和热情接待。感谢吉尔伯特·阿代尔、丽贝卡·高尔斯、埃米莉·高尔斯、帕特里克·高尔斯、乔书亚·卡茨和埃德蒙·托马斯阅读了本书的初稿。他们非常聪明，知识丰富，实在不能算作普通读者，不过还是能够让我放心，至少某些非数学专家是能够读懂我的作品的。基于他们对此书的评论，我作出了许多改进。我把这本书献给埃米莉，希望她能够借此了解一点点我整天都在做的是些什么事情。

【注释】

[1] 约相当于初中水平。——译注，下同

[2] 分形的一种。

序言

李大潜

数学是绝大多数人学得最多的一门功课，但对于“数学是什么？”这一看来很普通的问题，却很难一下子给出一个使公众满意的回答。按照恩格斯的说法，数学是以现实世界的空间形式和数量关系为研究对象的。尽管人们现在对空间形式和数量关系的理解已经大大深化和拓展了，但作为一种哲学的概括，恩格斯的这一论断应该说并没有过时，也便于向公众表述并为公众所接受。然而，要真正深入地回答“数学是什么？”这个问题，不能仅仅从定义出发，而必须涉及数学的具体内涵，作一些比较深入的解释和说明，才能达到使人信服的效果。但是，要这样做，会常常碰到下面两个似乎难以克服的技术上的困难。

第一，数学内涵的展现离不开众多的术语、记号和公式。在公众对有关的数学内涵产生兴趣并开始有所领悟之前，很可能早已为这些术语、记号和公式搞得晕头转向甚至望而却步了。

第二，数学内涵的展现同样离不开必要的逻辑推理。推理若过分严密，很难引起公众的兴趣；但若过于粗疏，语焉不详，则又易使人不得要领。

在现在的这一本书中，看不到很多的术语、记号和公式，对有关的数学概念及内涵一律用简明而生动的语言来介绍，看似如数家珍，娓娓道来，但举重若轻，高屋建瓴，反而更好地揭示了本质。不熟悉有关数学内容的读者，会感到茅塞顿开、豁然开朗；而已经熟悉有关内容的读者，也会有如沐春风、别开生面的感受。书中在论述极限时，用有限来刻画无限的生动而细致的处理方式，虽然本质上是经典的ε-N（或ε-δ）数学描述的一套直观而通俗的说法，但脱离了数学记号和公式，更显得清楚和亲切，就是一个典型的范例。书中对黄金分割比是一个无理数的证明、用地图册来介绍微分流形的概念等等，都同样独具匠心，可圈可点。另一方面，作者在不便于或不需要进行严格数学推导的时候，会巧妙地绕过去，但对于必要的推导和论证，绝不偷工减料，而是不遗余力地以十分详尽的方式加以说明，一步步地将读者引向有关的结论，同时也加深了读者对逻辑论证严密性的理解。在这方面，关于数系的扩张、关于$\sqrt{2}$是无理数的证明、关于平行公设的独立性等段落，都是倾心着力撰写的。正因为作者作了如此认真的努力，这一本篇幅不大的书显得出类拔萃，应该说为现有众多的数学科普读物提供了一个楷模。能够做到这一点，而且做得如此出色，不言而喻，归根结底是和作者深厚的数学功力、对数学内涵及其精神实质的深刻理解和把握分不开的。

本书一开始就讲了数学建模，指出数学研究的对象只是有关现实世界的数学模型，同时，指出有关的数学模型并不真正是相应的现实世界，而只是它的一个近似的代表与反映。在书中作者反复强调并解释的是他的一个基本的观点：对于数学，不要问它是什么，而只要问它能做什么。这一抽象化的思考方法，将重点放在数学内部体系的相容性，强调新的数学概念、方法与内容和已有的数学体系应自然地融为一体，强调要将有关的数学内容脱离其物理上的实在、变为符合一些特定规则的记号，就会更利于应用，更利于正确地理解高等的数学。作者在书中举

出指数函数、圆的面积、高维空间、分数维的几何等等一系列的例子来阐明这一观点对克服难关，深入理解与拓展数学概念所带来的好处。这是很有启发性的，也是很自然的，反映了抽象思维的必要性和优越性。由于有关的数学模型虽然只是现实世界的一个近似的代表，但毕竟是一个代表，适应于它的一些规则一开始并不是凭空而来的，而是从现实世界中移植、挪用或抽象过来的，对不同的现实世界可能引入不同的规则，也会造成不同的数学对象（不满足乘法交换律的矩阵，就是一个例子）。数学与现实世界的关系，套用一句文艺界的术语，看来应该是源于生活、高于生活的关系。如果作者在强调他的上述观点及做法的同时，也能够强调，数学要真正得到原创性的重大进展，除了要密切关注及面对数学内部的矛盾运动外，还要密切关注现实世界（包括其他科学技术）对数学提出的问题和需求，努力从外部世界中汲取生动活泼、丰富多彩的营养，应努力使二者相互促进、相得益彰，是不是会更全面、更富有启发性呢？

本书的作者蒂莫西·高尔斯（Timothy Gowers）教授是1998年获得菲尔茨（Fields）奖的著名数学家。2000年当我在法国巴黎访问时，因美国克雷（Clay）研究所给法国科学院院士阿兰·科纳（Alain Conne）教授颁发一个大奖，曾在法兰西学院（Collège de France）的讲演大厅里召开过颁奖会。会上获菲尔茨奖不久的蒂莫西·高尔斯教授应邀作了一个公众讲演。他在强调数学是一个整体的时候，曾说，如果把所有的数学分支按是否有联系组成一个网络，一定是一个连通的网络，而不会有一些学科，尽管它们看来与其他分支联系很少，游离于整个数学这一大网络之外。这正像有些人有很多亲戚朋友，有些则很少，但整个社会的人群所组成的网络仍是连通的一样。他的这一观点及如此通俗易懂的说法曾给我留下了深刻的印象，从这个意义上说，我和他已有一面之缘。这次有机会看到他这一本颇具特色的数学科普著作的中译本问世，也是一件幸事，特为之序。

2013年12月25日

第一章
模型
扔石头问题

风轻云淡的一天，你站在水平地面上，手里拿着一块石头，想要扔得越远越好。已知你能用多大的力气扔出去，那么最重要的决策就是选择石头出手时与地面的夹角。如果夹角太小，那么尽管石头在水平方向的速度分量很大，也会很快落到地面上，因而飞不出太远；反之，若夹角过大，石头能在空中停留较久但掠过的水平距离却不远。很明显我们需要在这中间作一些权衡。

利用牛顿物理学和微积分的一些初步知识，可以计算得到最佳的折中方案——石头离手时应与地面呈45度夹角。就这个问题而言，这基本上是最简洁优美的答案了。同样的计算还可以告诉我们石头在空中的飞行轨迹是个抛物线，甚至还能得出脱手后在空中任意时刻的速度有多大。

看起来，科学与数学相结合能够使我们预测石块飞出去直至落地之前的一切行为。然而，只有在我们作了许多的简化假设之后才能够如此。其中最主要的假设是，作用在石头上的只有一种力，即地球的引力，而且这种力的大小及方向在各处总是一样的。但实际上并非如此，因为它忽略了空气阻力、地球自转，也没有计入月球的微弱引力，而且越到高处地球引力越小，在地球表面上“垂直向下”的方向也随着具体位置的不同而逐渐变化。即使你能够接受上述计算，45度角的结果也基于另一个隐含假设：石头离手的初始速度与夹角无关。这也是不正确的：实际上夹角越小，人能使上的力气越大。

上述这些缺陷的重要性各有不同，我们在计算和预测中应该采取怎样的态度来对待这些偏差呢？把所有因素全部考虑在内进行计算固然是一种办法，但还有一种远为明智的办法：首先决定你需要达到什么样的精确度，然后用尽可能简单的办法达到它。如果经验表明一项简化的假设只会对结果产生微不足道的影响，那就应当采取这样的假设。

例如，空气阻力的影响相对来说是比较小的，因为石头很小很硬，密度大。假如在出手角度上有较大的误差，那么通过计入空气阻力来将计算复杂化就没有多大意义。如果一定要考虑进去的话，以下这条经验法则就足矣：空气阻力变大，则通过减小出手角度来弥补。

何为数学模型

当我们考察一个物理问题的解答时，十有八九能够就其中科学贡献部分和数学贡献部分划出一道清晰的界线。科学家在观察和实验的基础上，作一些简洁性与解释有效性的一般性考虑，建立一种理论。数学家，或者做数学的科学家，则研究理论的纯粹逻辑结果。有时候，这些情形是常规计算的结果，常规计算所预言的现象正是理论在提出时所要解释的。在某些偶然的情况下，理论所作出的预言则完全出乎意料。如果这些意料之外的现象后来被实验所证实，那么我们就得到了支持这种理论的重要证据。

然而，由于我上面所讨论到的简化问题，“证实一项科学预言”的概念就多多少少有了些问题。让我们考虑另一个例子：牛顿的运动定律和引力定律告诉我们，两个物体从同样的高度开始作自由落体运动，它们将同时到达地面（如果地面平坦）。这种现象由伽利略首先提出，它有点违背我们的直觉。实际上，它违背的不仅是我们的直觉：如果你亲自试一试，比方说用高尔夫球和乒乓球，你会发现高尔夫球首先落地。既然如此，究竟在什么意义上伽利略的论断是正确的呢？

当然，由于空气阻力的存在，我们不可能把这个小实验当作伽利略理论的反例：实验证明，当空气阻力很小时理论是正确的。如果你对此有所怀疑，觉得空气阻力实在稀松平常，怎能总是挽救牛顿力学的预测于失败之际，那么，找个机会去观察一下羽毛在真空中的下落，你就能重拾对科学的信念以及对伽利略的赞赏——真空中，羽毛的下落的确与石头别无二致。

尽管如此，由于科学观察永远不是完全直接性和决定性的，我们仍需要一种更好的方式来描述科学与数学之间的关系。数学家并不是将科学理论直接应用于现实世界中，而是应用于模型上。在这里，模型可以看作是所要研究的那部分现实世界的一种虚构、简化的版本。在模型里，我们就有可能进行完全精确的计算。在扔石头的例子中，现实世界与模型的关系正如同图1和图2所展示的那样。

对于一种给定的物理情形，有多种方法将其模型化。我们需要结合切近的经验与深入的理论考量来决定，哪种模型更有可能向我们透露世界的本真。选择模型时，有一个需要优先考虑的因素，即模型的行为应当与实际中观察到的行为密切对应。


图1 飞行中的球甲

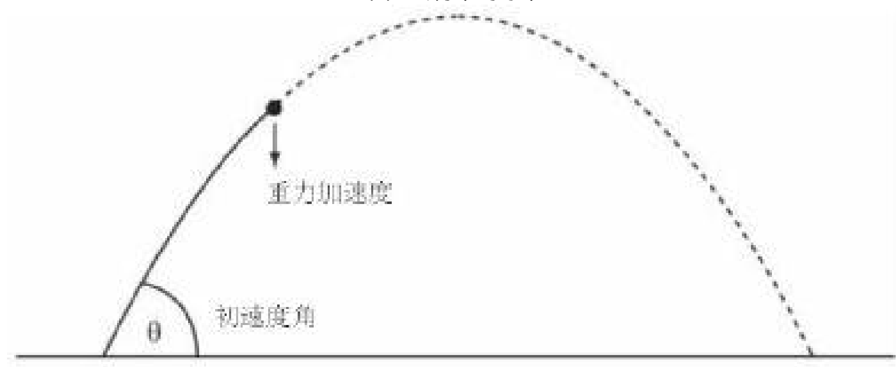


图2 飞行中的球乙

但是，诸如简洁、数学表达上的优雅等其他因素可能反而时常会更重要一些。实际上，确实有一些模型在现实世界中几乎找不到任何相似之物，但同时却非常有用。接下来有一些例子将会对此进行说明。

掷骰子问题

假设我要掷一对骰子，想要了解它们的行为如何。经验告诉我，问某些问题根本是不现实的。例如，不可能期待有人能预先告诉我某一次掷骰子的结果，即便是他掌握了很高超的科技，并且用机器来掷骰子。与此相反的是，关于概率的问题则常常是能够回答的，比如“两个骰子的结果之和为7的可能性有多大”之类的问题。这样的问题的答案可能也是有用处的，比方说在我玩双陆棋[1]赌钱的时候。这一类问题很容易模型化，只要将两个骰子掷出来的结果看作是从下列36个整数对中随机选取一个。

（1,1）	（1,2）	（1,3）	（1,4）	（1,5）	（1,6）
（2,1）	（2,2）	（2,3）	（2,4）	（2,5）	（2,6）
（3,1）	（3,2）	（3,3）	（3,4）	（3,5）	（3,6）
（4,1）	（4,2）	（4,3）	（4,4）	（4,5）	（4,6）
（5,1）	（5,2）	（5,3）	（5,4）	（5,5）	（5,6）
（6,1）	（6,2）	（6,3）	（6,4）	（6,5）	（6,6）

每组数对中第一个数表示骰子甲的结果，第二个数表示骰子乙的结果。恰有六组满足两数之和为7，因此掷出7的概率就是6/36，即1/6。

可能有人会反对这种模型，他们会说，骰子在滚动时是遵循牛顿定律的，至少在很高的精度上遵循，因此骰子落地的情况根本不是随机的：原则上是完全能够被计算出来的。但是，“原则上”这个短语在这里被过度使用了，因为这样的计算将会是

极端复杂的，并且需要知道骰子的形状、材料、初始速度、旋转速度等更为精确的信息，而这般精确的信息在实际中是根本无法测出来的。因为这一点，使用某种更为复杂的决定论模型是无论如何也不会有任何优势的。

预测人口增长

较“软”性的科学——比如生物学和经济学中，也充斥着各种数学模型，这些模型都远比它们所要真正表示的现象简单得多，甚至以某些方式刻意地使其不够精确，但是这些模型还是有其用场、富于启发性。就以一个在经济学上有重要意义的生物学问题为例，我们来考虑预测一国未来20年的人口。我们可能会用到一种非常简单的模型，即将全国人口表示为一组数对 $(t, p(t))$，其中 t 表示时间，$p(t)$ 表示时刻 t 的人口规模。另外我们要用到两个数 b 和 d，来表示出生率和死亡率。所谓出生率和死亡率，即每年出生人数和死亡人数占总人口的比例。

假设我们已知2002年初的总人口是 p。根据上述模型，2002年的出生人数和死亡人数将分别为 bp 和 dp，因此2003年初的总人口将为 $p+bp-dp=(1+b-d)p$。其他年份亦然，因此我们就能够写出公式 $p(n+1)=(1+b-d)p(n)$，意即 $n+1$ 年年初的人口是 n 年年初人口乘以 $(1+b-d)$。换句话说，每一年人口数量都会乘上 $(1+b-d)$。那么20年后的人口就是乘以 $(1+b-d)^{20}$，于是就得出了初始问题的答案。

这个模型已经比较好了，它能向我们证明，如果出生率明显高于死亡率，那么人口就会急剧增长。但即便如此，它也还是不够现实的，它作出的预测可能很不精确。比方说，模型中假设出生率和死亡率在20年中都保持不变，这并不太可信。过去的事实已经证明，出生率和死亡率经常会受到社会变迁和政治事件的影响，如医学进步、新型疾病出现、女性首次生育年龄增大、税负激励以及偶尔发生的大规模战争等等。生育率与死亡率会随时间变化还有另一个原因，就是一国国民的年龄分布可能相当失衡。比方说，15年前出现了一波婴儿潮，那么我们就有理由预期再过10年到15年出生率就会增加。

因此，通过引入其他因素来使模型复杂化，这个想法相当诱人。我们记出生率和死亡率分别为 $b(t)$ 和 $d(t)$，使其可以随时间变化。我们并不想单用一个数字 $p(t)$ 来表示总人口，我们可能想要知道不同年龄层各有多少人。如果同时还能尽可能多地知道各个年龄层的社会态度和行为倾向，也会对预测未来的出生率和死亡率有所帮助。获取这样的统计信息是十分昂贵且困难的，但这些信息确实能够大幅提高预测的精度。因此，没有一种模型能够脱颖而出，声称比其他模型都好。关于社会和政治的变迁，谁也不可能确切地说出情况会是什么样子。关于某种模型，我们所能提出的合理问题，最多只能是问某种有条件的预测，也就是说模型只能告诉我们，这样的社会或政治变迁如果发生的话会产生怎样的影响。

气体的行为

气体动理论由丹尼尔·伯努利在1738年提出，后来又由麦克斯韦、玻尔兹曼等人在19世纪后半叶推进。根据这种理论，气体是由运动着的分子组成的，气体的许多性质——如温度和压强，都是这些分子的统计属性。譬如，温度就对应着分子的平均速度[2]。

有了这样的想法之后，让我们设想一种模型来描述方盒子中的气体。这个盒子当然应该用一个立方体来表示（意即数学的而非物理的）。既然分子是非常小的，那么用立方体中的点来表示也就很自然了。这些点应当是运动的，所以我们必须确定控制它们运动的规则。此时，我们需要作出一些选择。

如果盒中只有一个分子，那么规则可以很明显：分子以恒定速度运动，撞到盒子壁面时就反弹出去。要将这种模型推广到包含N个分子的情形（N是个较大的数），最简单的办法就是假设分子都遵从这样的运动规则，分子之间绝对没有相互作用。为了启动这样的N分子模型，我们要选择分子（或者说，表示它们的那些点）的初始位置及初始速度。随机选择是一种好办法，因为我们可以预期，在任意时刻，真实气体中的分子都在空间中弥散着，运动方向也各式各样。

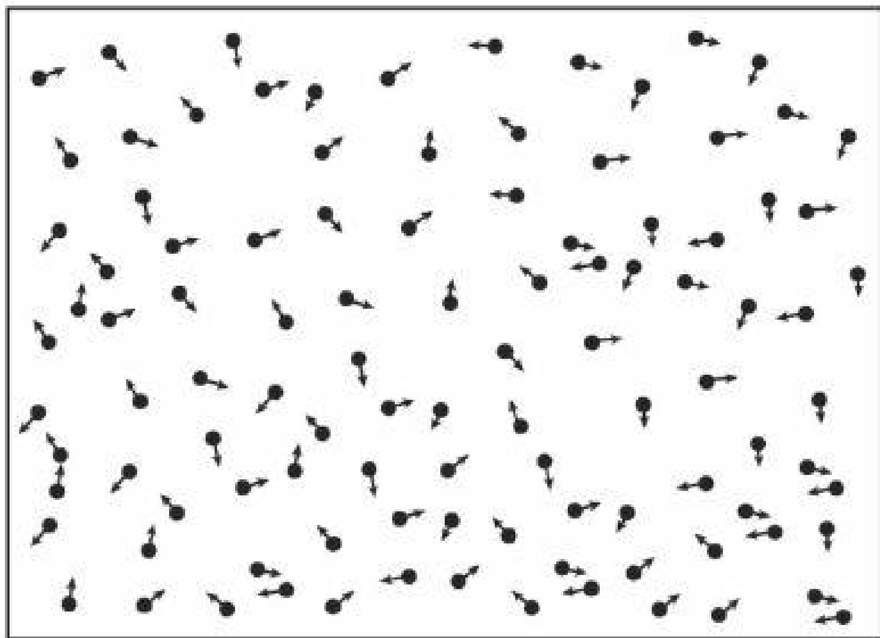


图3 气体的二维模型

要说清在立方体中随机取一个点并不困难，随机的运动方向也不复杂，但如何随机地选择速率就有些含混了，因为速率可以取从0到无穷大的任意值。为了避免这个困难，我们可以作一个从物理角度看似不太可信的假设，让所有分子的速度大小都相同，仅仅让初始位置和方向能够随机选取。图3就表示了这个模型的一个二维情形。

N个分子完全相互独立运动的假设毫无疑问是过度简化的。比方说，利用这个模型，我们就不可能理解，为什么当温度足够低时气体会液化：当你把模型中的各点运动速率降低，得到的还是相同的模型，无非

跑得慢一些而已。不过这个模型还是能够解释真实气体的许多行为。例如，想象盒子被慢慢压缩的情形。分子仍然会继续以相同速率运动，但由于盒子变小，分子撞击壁面更加频繁，可供撞击的壁面面积也变小了。由于这两个缘故，单位面积的壁面每秒钟被撞击次数就增多了。这些撞击正是气体压强的来源，于是我们可以总结出，气体体积减小时，气体压强很可能增大——正如实际观测所证实的那样。类似的论证还可以解释，为什么气体温度升高而体积不变时，压强会增大。要推算出压强、温度与体积之间的数值关系也并不困难。

上述模型大致上就是伯努利所提出的模型。麦克斯韦的成就之一就是发现了一个优美的理论，来解决如何更逼真地选择初始速率的问题。为了理解这一点，让我们放弃分子间没有相互作用的假设。作为替代，我们假定分子会时不时地相互碰撞，就像台球一样。碰撞之后，它们就以另外的速率、向另外的方向，在遵守能量守恒和动量守恒定律的前提下随机弹开。当然，既然我们用没有体积的点来表示分子，那么就很难看出它们要如何碰撞。不过，这个麻烦在理论中恰可以作为一个非正式的论据，说明分子运动速度及方向具有某种随机性。麦克斯韦就这种随机性的本质作了两个非常合理的假定：其一，分子运动的随机性不随时间而改变；其二，这个随机性在不同方向上没有区别。大体来讲，第二条假设意味着，选取d_1和d_2两个方向及某个速率s，那么粒子以速率s沿着d_1方向运动的概率和以速率s沿着d_2方向运动的概率是相同的。不可思议的是，这样的两个假设就足以恰好决定分子运动速度的分布形式，即意味着，如果我们想要随机选取速度，就只有一种自然的方式。（它们应当服从正态分布。这种分布产生了著名的“钟形曲线”。这种曲线在各种各样的场合下经常出现，既出现在数学中也出现在实验中。）

一旦选定了速度，我们就可以再次忘掉分子间的相互作用。结果表明，这种作了一点改进的模型依然存在着原始模型中的许多瑕疵。为了进一步修正，我们只能再把分子间的相互作用考虑进来。但是结果发现，即使是非常简单的相互作用粒子模型，其行为也极其复杂，会引发极为难解，事实上多数都未能解决的数学问题。

大脑和计算机的模型化

计算机同样也可以看作由相互作用的各部分集合而成；很大程度上由于这个原因，理论计算机科学中同样有很多悬而未决的重要问题。其中有如下这样一个例子，我们可能愿意去尝试解答。假设某人选取了两个素数p和q，将它们相乘后的结果pq告诉你。你只要逐个取素数，看看它是否能够整除pq，即可以找出p和q。例如，给出91，你很快就能发现它不是2，3，5的倍数，继而发现它恰好等于7×13。

然而，当p和q非常大时——比方说都是200位的素数，那么这一试错过程会耗时极长，即使借助于强力计算机也是如此。（如果你想要体会一下这种困难，不妨尝试找出6901的两个素因子，以及280 123的。）可另一方面，似乎也不难感觉到，说不定这个问题存在着更聪明的解决办法，基于它就可以编制出一种快速运转的计算机程序。如果能找到这种好办法，我们就能破解作为大部分现代安全系统之基石的密码，包括在互联网上以及其他各处——破解这些密码的难点就在大整数的因子分解。反之，如果能够表明由pq计算出p和q的这种快速有效的方法不存在的话，我们则能够安心。不幸的是，虽然计算机总在不断地让我们惊叹它的各种能力，对于它们做不了的，我们却几乎毫无了解。

在思考这个问题之前，我们必须找到一种方法来数学化地表示计算机，并且要尽可能简单。图4所显示的就是一种极好的方法。它包含许多层节点，连结点和点之间的线段称作“边”。进入到最顶层的称作“输入”，这是一条0和1的序列，从最底层出来的叫作“输出”，是另一条0和1的序列。节点分为三种，分别称作“与门”、“或门”和“非门”。每一个门都从连结上层的边中接收到一些0和1。它再根据所接收到的数码来自己发出一些0和1，所遵循的简单规则如下：与门当接收到的输入全部为1时，输出1，否则输出0；或门当接收到的输入全部为0时，输出0，否则输出1；非门只允许一条边连结上层，它在接收到1时输出0，接收到0时输出1。

一系列门由边连接起来就称作一条电路，我上面所描述的模型正是关于计算的电路模型。使用“计算”一词是恰如其分的，因为我们可以把电路看作这样一种装置，它拥有一条0和1的序列，继而按某些预定规则将其变换为另一条序列，如果电路很大，变换规则可能会很复杂。这也正是计算机所进行的工作，只不过它们能够把这些序列翻译成我们能够理解的格式，诸如高级程序设计语言、视窗系统、图标等等。实际上，存在一种比较简单的办法（仅从理论的角度而言——在实践上操作将是个噩梦），能将任意计算机程序转换为一条按完全相同的规则变换01序列的电路。而且，计算机程序的重要特征恰在其对应电路中有着非常类似的对照物。

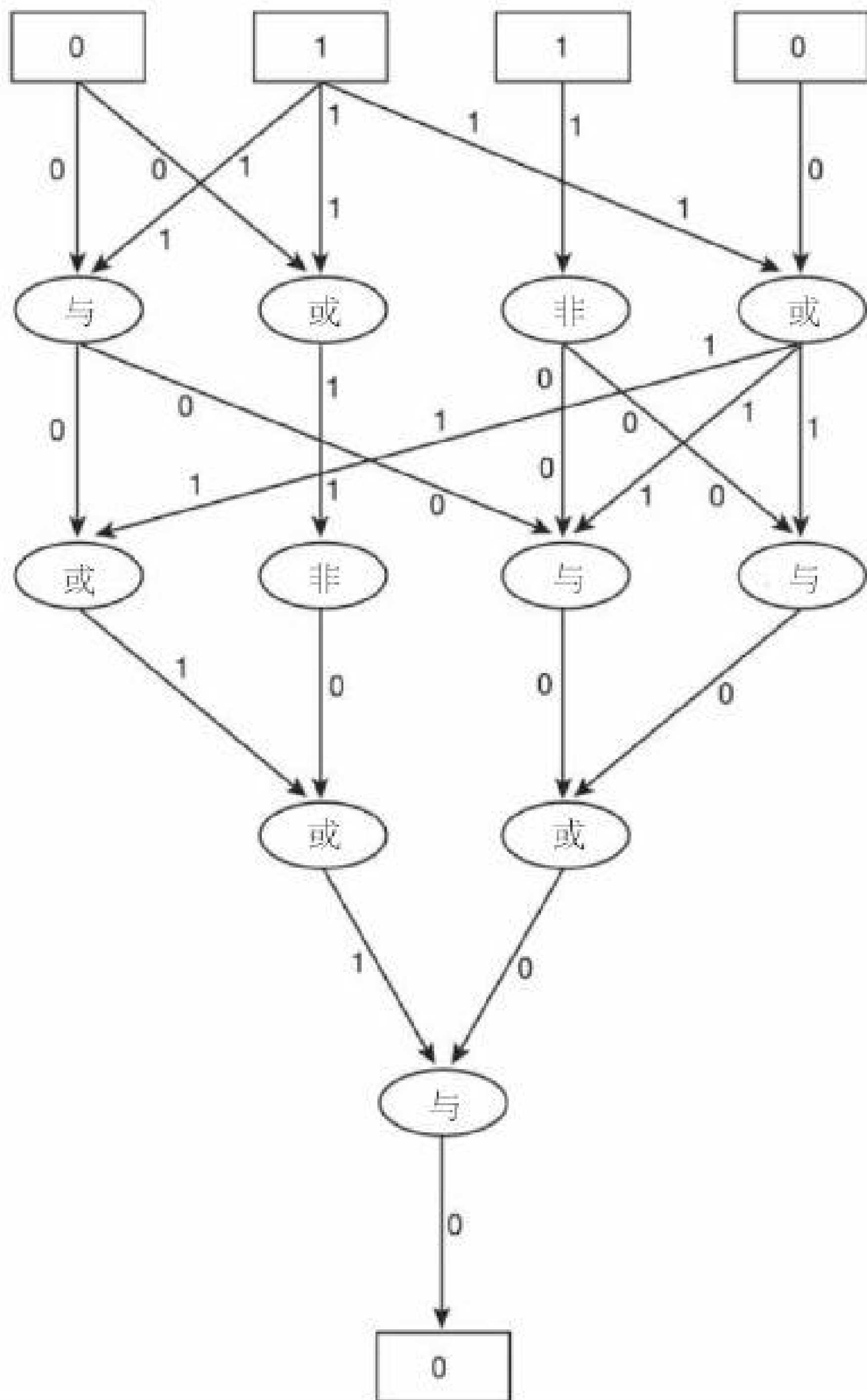


图4 一个简单的计算机程序

具体而言，电路中的节点数量正对应于计算机程序运转所花费的时间。因此，如果我们能够表明，按某种方式来变换01序列需要庞大的电路，那么也就说明这种变换方式所需要的计算机程序运转时间很长。我

们使用电路模型而非直接分析计算机，其优势就在于，从数学的角度来看电路更简单也更自然，考虑起来更容易一些。

对电路模型进行一点小小的修改，我们就能得到大脑的一种有用的模型。这种模型不再使用01序列，而是使用0和1之间的任意值来表示强度各异的信号。所有的门，即对应于神经元或者脑细胞，也有所不同，但其行为还是很简单的。每个门从其他的门接收到一些信号，如果这些信号的总强度——对应数字的总和——足够大，门就在某个特定的强度水平上输出它自己的信号，否则不输出。这对应于神经元所作的是否“激发”的决策。

似乎很难相信这个模型能够捕捉大脑全部的复杂性，但这部分缘于我并没有提到应当有多少个门以及如何安排这些门。一个典型的人类大脑包含大约1000亿个神经元，它们以非常复杂的方式排列着；以我们当前对大脑的认识，还不可能谈及太多——至少在精微的细节方面不可能说清。不过，上述模型提供了一种有益的理论框架，供我们思考大脑可能是如何工作的，也使我们能够模拟某些类似于大脑运行的行为。

地图染色与时间表制定

设想你正在绘制一幅地图，地图上分成了若干区域，你希望为这些区域选取颜色。你可能想选用尽可能少的颜色，但同时还希望避免任意两块相邻区域使用相同的颜色。再设想你正在安排大学课程的时间表。课程有很多门，但可供安排的总时间段有限，所以会有某些门课程时间冲突。哪些学生选了哪些课程已经登记在列，你希望尽可能合理安排，仅当两门课程没有学生同时选择时才可以时间冲突。

这两个问题看似截然不同，但一种合理的模型能够说明，从数学的观点来看它们其实是一样的。在这两个问题中，都需要给一些对象（国家、课程）赋予一些属性（颜色、时间）。对象中有某些两两组合（相邻的国家，不能冲突的课程）是不能相容的，也就是说它们不能被赋予相同的属性。在这两个问题中，我们其实并不关心具体的对象是什么、要赋予的属性是什么，所以我们也可以仅用点来表示它们。为了表示那些不相容的成对的点，我们可以将它们用线段连结起来。这样一组边和边连结起来的点的集合，就是“图”这种数学结构。图5给出了一个简单的例子。通常称图中的点为顶点，称线段为边。

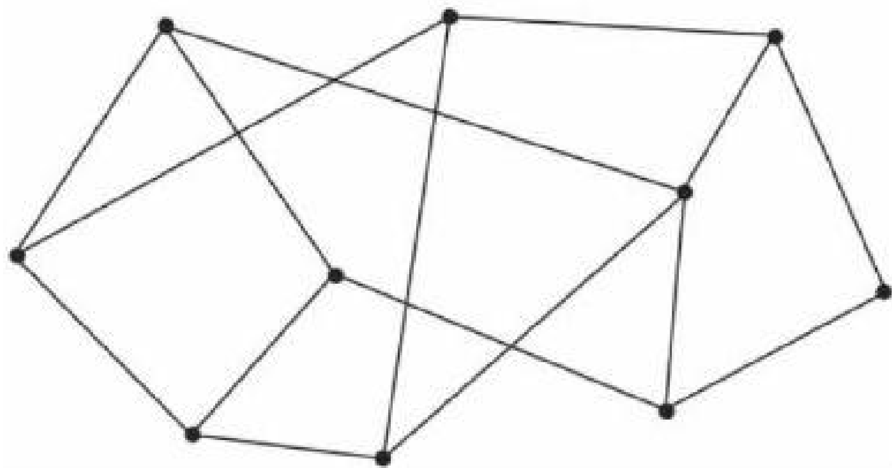


图5 10个顶点和15条边组成的图

一旦我们将问题用这种形式表示出来，我们在两个例子中的任务就统一为：将顶点分成尽可能少的几组，使得每组中不包含由同一条边相连的两顶点。（图5所表示的图可以分成三组，但不可能分成两组。）这也就说明了使模型尽可能简化的另一个充足理由：如果幸运的话，同样的模型可以用来一次性研究很多不同的现象。

“抽象”一词的不同含义

在设计模型的时候，我们会忽视所考察的现象中尽可能多的信息，从中仅仅抽象出那些对理解其行为必不可少的特征。在上述我所讨论的例子中，石头被简化为一个点，一国的全部人口被简化为一个数，大脑被简化为遵循一些简单数学规则的门的网络，分子间相互作用被简化到根本不存在。结果得到的数学结构就是具体情形在模型化之后的抽象表示。

我们说数学是一个抽象的领域，这包含两层含义：一来它从问题中抽象出重要特征，二来它所处理的对象不是具体的、有形的。在下一章，我们将讨论数学抽象的第三层，也是更深层的含义，前面的例子其实已经让我们对此有所了解。图是一种具有可塑性的模型，可以用在多种场合。但当我们研究图时，完全不需要考虑它的这些具体用途：点究竟表示地区、课程还是别的完全不同的东西，这并不重要。研究图的理论工作者可以完全抛弃现实世界，进入到纯粹抽象的王国之中。

【注释】

[1] 一种棋子游戏，靠掷骰子决定行棋格数。

[2] 原文如此，有误，应为“平均动能”。

第二章
数与抽象
抽象方法

几年前，《泰晤士报文学增刊》的一篇评论在开篇写到：

> 已知$0\times 0=0$以及$1\times 1=1$，就可以得到：平方等于自身的数是存在的。但是再进一步，我们就可以得到：数是存在的。经过这简简单单朴实无华的一步，我们似乎就从一个基本的算术式得到了一个令人吃惊的、极具争议的哲学结论：数是存在的。你可能还以为这有多么困难呢。
>
> A.W.莫尔评论杰罗尔德·J.卡茨《现实理性主义》，
> 《泰晤士报文学增刊》，1998年9月11日

对于这个论证，我们可以从许多角度来批评，而且基本上不可能有人把它当真，包括评论人自己。然而，确实有哲学家在严肃地思考数是否存在这样的问题，而且正是这种思考使得他们不同于数学家——数学家理所当然地认为数就是存在的，不理解它何以能够成为一个问题。本章的主要目的就是要解释，为什么数学家能够愉快地忽视这样一个看似非常基本的问题，甚至理应忽视它。

如果我们把同样的论证方法搬到国际象棋上来，这个“简简单单朴实无华”的关于数的存在的论证就明显荒谬了。已知象棋中的黑色国王有时可以斜向移动一格，由此得出，存在着可以斜向移动一格的棋子。但接下来将进而得出：国际象棋的棋子是存在的。当然了，我想说的并不是那句平平常常的断言：人们有时候会制造国际象棋的棋具。毕竟，没有棋具同样也能下棋。我所指的是一个更加“惊人”的哲学结论：象棋棋子的存在独立于它们的物理形态。

在象棋中，黑色国王是什么呢？这是个奇怪的问题。最令人满意的处理方式似乎是，回避这个问题。除了指着棋盘，说明游戏的规则，我们还能做些什么呢？也许在这么做的同时会对黑色国王给予特别的关注？真正重要的问题并不是黑色国王的存在性或者它的本质属性，而是它在游戏中所发挥的作用。

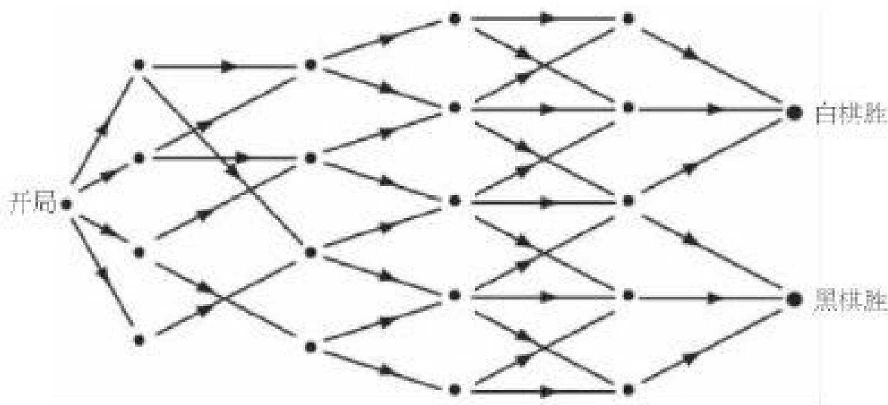
人们所谓的数学中的抽象方法，正是我们采取类似态度来对待数学对象的结果。这种态度能够用这样一句话涵盖：数学对象是其所做。在语言哲学中，我们能发现类似的话经常出现，而且可能饱受争议。我们可以举出两个例子，比如索绪尔所说的“语言中只有差异”，还有维特根斯坦所说的“语词的意义是它在语言中的用法”（参见书末“延伸阅读”）。另外，我们还可以加上逻辑实证主义者的宣言：“陈述的意义就是其证实的方法。”如果你出于哲学方面的考虑，认为我的观点令人生厌，那你不妨不要将它看作一个教条式的断言，而是视为一种可供采纳的态度。实际上，我希望表明的是，要想正确地理解更高等的数学，采纳这种态度是至关重要的。

没有棋子的象棋

看出这一点很有意思——尽管我的论述并不直接依赖于它：象棋，或者任何类似的游戏，都可以以图为模型。（图已经在前一章末定义过了。）图的顶点代表游戏的某种可能的局面。如果两个顶点P和Q有边相连，那就意味着可以从局面P出发，经过合乎规则的一步之后达到局面Q。因为有可能无法从Q返回到P，所以这样的边需要用箭头来指示方向。某些顶点可以看作白棋获胜，还有某些顶点可以看作黑棋获胜。游戏从一个特定顶点，即游戏的开始局面出发。两位棋手相继沿着边移动。第一位棋手努力要走到白棋获胜的某个顶点，第二位棋手则要走到黑棋获胜的顶点。图6显示了某种大大简化的类似游戏。（图中不难看到，对于这个游戏来说白棋有必胜策略。）

尽管象棋的这种图论模型很难有应用上的意义，因为现实中可能的局面数量实在太庞大了，但是就其所表示的游戏和象棋完全等价来说，它仍然是一种完美的模型。不过在定义它的时候，我完全没有提及关于棋子的任何事情。从这个角度来看，黑色国王是否存在的问题就是个离奇的问题了：棋盘和棋子只不过是方便我们将这么大的图中一团乱麻似的顶点和边组织起来的一种原则而已。如果我们说出“黑色国王被将军了”这样的句子，那么这只是一种简化的说法，它所指的无非是两位棋手到达了极多的顶点中的某一个。

图6 白先，白方有必胜策略



自然数

“自然”是数学家对我们所熟悉的1，2，3，4这样的数字所赋予的称呼。自然数是最基本的数学对象，但它们似乎并没有引导我们去抽象地思考。毕竟，单单一个数字5能做些什么呢？它不可能像个棋子一样走来走去。它所具有的似乎是一种内在的属性——某种纯粹的“五性”，当我们观察图7这样的图片时就能立即从中提炼出这样的性质。

然而，当我们考虑更大的数时，其中的纯粹性就变少了。图8向我们表示了7，12和47这几个数。可能有部分人能够立刻把握第一张图中的“七性”，但大多数人可能会在一瞬间有这样的思考：“外围的点形成了一个六边形，再加上中心的一个得到6＋1＝7。”类似地，12可能会被考虑为3×4或2×6。至于47，和别的数字比起来，比方和46相比，一组这个数量的物体就缺乏特别之处。如果这组物体以某种模式排列起来，例如排成少两个点的7×7的阵列，那么我们的知识就可以通过7×7-2＝49-2＝47迅速地得出总数一共有多少。如若不然，我们就只能去一个个地数，这时我们则是将47视为46之后的那个数，而46又是45之后的那个数，依此类推。



图7 “五性”的概念

也就是说，当数字变得还不算太大时，我们就已经不再将其视作一些独立的客体了，而开始通过它们的内在属性，它们与其他数字的关联，以及它们在数系中的作用来理解。这也就是我之前说数能“做”什么所要表达的意思。

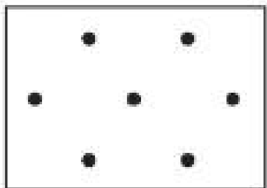
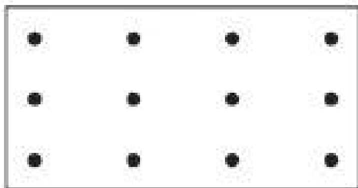
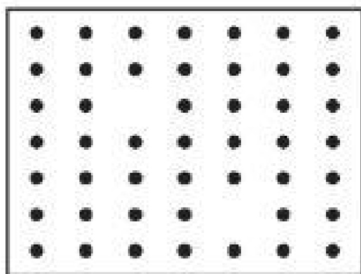
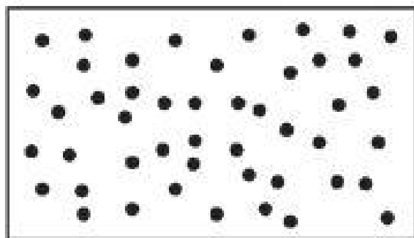





图8 表示7，12和47（两种）的方式

如我们已经清楚地看到的，数的概念与加法、乘法这样的算术运算紧密相连。比方说，如果没有算术的概念，对1000000017这样的数的意义就只能有很模糊的把握。一个数系并不仅仅是一堆数字，而是由数字及算术规则共同构成的。我们还可以这样来总结这种抽象方法：考虑规则，而不是考虑数字本身。按这种观点，数字就可以被当作某种游戏中的记号（或许应该被称为计数子）。

为了对其中的规则有一些了解，让我们来考察一个简单的算术问题：如果我们要确证38×263＝9994，应当做什么？多数人可能会用计算器来检查，但要是因故无法这样检查呢？他们可能会作如下的推理。

$$\begin{aligned}38\times263&=30\times263+8\times263\\&=30\times200+30\times60+30\times3+8\times200+8\times60+8\times3\\&=6000+1800+90+1600+480+24\\&=9400+570+24\\&=9994\end{aligned}$$

为什么上述几步看起来是如此地天经地义呢？比方说，为什么我们会不假思索地相信30×200＝6000？30的定义是3×10，200的定义是2×（10×10），所以我们可以充分相信30×200＝（3×10）×（2×（10×10））。但为什么是6000呢？

一般没人会去问这种问题，不过既然有人问了，我们可能会说：

$$(3\times10)\times(2\times(10\times10))=(3\times2)\times(10\times10\times10)=6\times1000=6000$$

我们并没有经过仔细考虑就利用了关于乘法的两个熟知事实：其一，两数相乘时谁先谁后并没有关系；其二，多个数相乘时无论怎样加括号都没有区别。例如，7×8＝8×7以及（31×34）

×35＝31×（34×35）。第二个例子中，中间结果肯定会受到括号位置的影响，但我们知道最终结果是相同的。

这两条规则被称为乘法的交换律和结合律。下面我将列出若干条加法和乘法中常用的规则，也包括上述两条。

A1 加法交换律：对任意两个数a和b，有a＋b＝b＋a。

A2 加法结合律：对任意三个数a、b和c，有a＋（b＋c）＝（a＋b）＋c。

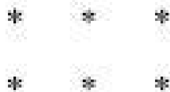
M1 乘法交换律：对任意两个数a和b，有ab＝ba。

M2 乘法结合律：对任意三个数a、b和c，有a（bc）＝（ab）c。

M3 1是乘法单位元：对任意数a，有1a＝a。

D 分配律：对任意三个数a、b和c，有（a＋b）c＝ac＋bc。

我列出这些规则并非试图告诉你这些规则本身多么有意思，而是想请你关注它们在我们的思考中所扮演的角色——即便是在相当简单的数学陈述中。我们对3×2＝6的信心大概是基于这样的一种图像：



这样一来，要直接论证38×263＝9994就是不可能的了，于是我们要以完全不同的方式来思考这一稍显复杂的事实，其中就要利用到交换律、结合律和分配律。如果的确遵守了这些规则，我们就会相信最后的结果。而且，虽然绝不可能对9994个物体有视觉上的感知，我们也相信结果是正确的。

零

在历史上，数字0的思想的诞生晚于正整数。这在很多人看来是个神秘而且矛盾的概念，它引发了类似这样的问题："某种事物既然没有，怎么可能存在呢？"但从抽象的观点来看，0其实很明确：它只不过是引入到我们数系中的一个新记号而已，并且满足下面这条特殊的性质。

A3 0是加法单位元：对任意数a，有0＋a＝a。

这就是关于0你所需要知道的一切了。无关它意味着什么，只是一条小规则告诉你它做什么。

那关于数字0的其他性质呢，比方说0乘以任何数都等于0的性质？我并没有列出这条规则，因为我们可以从性质**A3**及之前的其他规则把它推导出来。例如，我们已经将2定义为1＋1，那要怎样来表明0×2＝0呢？首先，根据规则**M1**有0×2＝2×0。然后，由规则**D**得到（1＋1）×0＝1×0＋1×0。但根据规则**M3**，1×0＝0，所以该式等于0＋0。规则**A3**意味着0＋0＝0，于是我们的论证就完成了。

从非抽象的角度出发，可能会这样去论证：0×2的意思是指，总共**0**个**2**相加，没有2，就是0。但用这种思考方式，我们将不太容易回答诸如我儿子约翰问我的这个问题（在他六岁时）：既然无和无相乘的意思是没有无，那为什么结果又会是无呢？尽管当时可能不太适合，但一个好的回答终归是，它能够如下所述从基本规则中推导出来。（每一步之后我都列出了所用到的规则。）

$0=1\times 0$	**M3**
$=(0+1)\times 0$	**A3**
$=0\times 0+1\times 0$	**D**
$=0\times 0+0$	**M3**
$=0+0\times 0$	**A1**
$=0\times 0$	**A3**

为什么我要对非常基本的事实给出如此冗长的证明呢？和上面一样，原因并不是我觉得这些证明多么有数学趣味，而是想表明，抽象地（利用几条简单规则，忽略数字的具体意义）而非具体地（考察数学陈述的实际意义）证明算术陈述是怎么一回事。将实际意义及思维图像与数学对象结合起来固然非常有用，但是，正如我们将多次在本书中看到的，这样的结合常常并不足以告诉我们在新的不熟悉的场合下应当怎样去处理。因而，抽象的方法是不可或缺的。

负数和分数

谁如果曾经给小孩子教过数学，那一定会知道，减法和除法并不像加法和乘法那样直接，他们学习起来更困难一些。为了解释减法，我们当然可以利用“拿走”的概念，提这一类的问题：“一开始有5个橘子，吃掉2个，还剩下几个？”然而这并不总是思考减法的最佳方式。例如，当我们用100减98时，更好的想法不是从100中取走98，而是考虑什么数加上98能够得到100。也就是说，更有效的做法是求解方程$98+x=100$，尽管计算时字母x并不常出现在我们的脑子里。类似地，考虑除法也有两种方式。为了解释50除以10的意义，我们既可以问“50个东西分成相等的10组，每组几个”，也可以问“10个东西分一组，50个东西能分几组”。第二种办法等于在问“10和几相乘能够得到50”，也就等于解方程$10x=50$。

向小孩解释减法和除法还有着更进一步的困难，那就是这两种计算并非总能够进行。例如，不可能从装有7只橘子的碗中拿走10只，11颗弹珠不可能平均分给3个小孩。但成人却能够计算7减去10和11除以3，分别得到-3和11/3。问题随之而来：-3和11/3这样的数实际上存在吗？如果存在它们又是什么呢？

从抽象的角度看，处理这个问题的方式类似于之前对于零的处理方式：全都抛至脑后。关于-3，我们只需要知道它和3相加等于0即可；关于11/3，只需知道它乘以3等于11即可。这就是它们的运算规则。再与之前的规则相结合，我们就可以在更大的数系中进行算术运算。为什么我们希望照这样扩充数系呢？原因就是，这样得到的模型允许我们在其中求解$x+a=b$和$ax=b$这样的方程，无论a和b取何值——只要a在第二个方程中不为0。换句话说，这样得到的模型中，减法和除法总是能够进行的，只要除数不为0。（除数为0的问题我们会在本章稍后谈到。）

按照这样的路数，我们只需再增加两条规则来扩充我们的数系：一条给我们带来负数，另一条给我们带来分数，即我们所熟知的有理数。

A4 加法逆元：对任意数a，总存在一个数b使得$a+b=0$。

M4 乘法逆元：对任意不为0的数a，总存在一个数c使得 $ac=1$。

确定了这些规则后，我们就可以将-a和1/a分别看作**A4**中的b、**M4**中的c的代号。至于更一般的表达式p q，它表示p乘以1q。

规则**A4**和**M4**还蕴含了另外两条规则，即消去法则。

A5 加法消去律：对任意三个数a、b和c，若$a+b=a+c$，则$b=c$。

M5 乘法消去律：对任意三个数a、b和c，若$ab=ac$且a不为0，则$b=c$。

第一条可以通过在等式两端加上-a得到证明，第二条可以通过在等式两端乘以1/a得到证明。应当注意**A5**和**M5**的地位与之前的那些规则是不同的——这两条是之前规则的推论，而不是我们向游戏中引入的新规则。

如果有人要加两个分数，如2/5加3/7，那么常用的方法是找出它们的公分母：

$$\frac{2}{5}+\frac{3}{7}=\frac{14}{35}+\frac{15}{35}=\frac{29}{35}$$

这种方法及类似方法的合理性可以用我们的新规则来证明。例如：

$$35\times\frac{14}{35}=35\times\left(14\times\frac{1}{35}\right)=(14\times 35)\times\frac{1}{35}=14\times\left(35\times\frac{1}{35}\right)$$
$$=14\times 1=14$$

又有

$$35\times\frac{2}{5}=(5\times 7)\times\left(2\times\frac{1}{5}\right)=(7\times 5)\times\left(\frac{1}{5}\times 2\right)$$

$$=\left(7\times\left(5\times\frac{1}{5}\right)\right)\times 2=(7\times 1)\times 2=7\times 2=14$$

于是，由规则**M5**得到2/5和14/35是相等的，正如我们在计算中所假设的那样。

类似地，我们还可以论证关于负数的一些熟悉的事实。请读者自己练习从上述规则中推出（-1）×（-1）＝1，它的推导和对0×0＝0的证明相当类似。

为什么在很多人看来，负数的实在性要低于正数呢？大概因为对数量不多的物体的计数是人类的基本活动，在这其中并不会用到负数。但这只不过意味着，作为模型的自然数系在某种特定场合下比较有用，而扩充数系则不太用得上。但如果我们考虑温度、日期或者银行账户，那负数就的确能够发挥作用了。只要扩充数系是逻辑自洽的——实际上也正是如此，用它作为模型就没有任何害处。

把自然数系称作一种模型似乎有点奇怪。难道我们不是在切实地数数，未涉及任何特定的理想化描述吗？我们的确是那样数数的，但数数的过程并非总是恰当的，甚至会根本不可能。从数学的观点来看，1 394 840 275 936 498 649 234 987这个数没有任何问题，但如果我们连佛罗里达州的选票都数不过来，就无法想象能确信自己拥有由1 394 840 275 936 498 649 234 987个东西组成的一堆。如果你将两堆落叶加到第三堆上，得到的结果并不是三堆树叶，而是一大堆树叶。倘若你刚观察过一场暴雨，那正如维特根斯坦所说，“‘你看到了多少水滴’这个

问题的恰当答案是，很多。并不是因为没有那么一个数字，而是因为你根本不知道有多少”。

实数和复数

实数系包含了所有能够用十进制无穷小数表示的数字。这个概念看似简单，实则不易，其中的缘由我们会在第四章加以解释。而现在，让我们来讨论一下将有理数系扩充到实数系的原因。我要讲的是，这原因正与引入负数和分数的理由类似：它们使我们能够求解某些方程，缺少它们我们则无法求解。

这些方程中最著名的一个莫过于$x^2=2$。在公元前6世纪，毕达哥拉斯学派发现$\sqrt{2}$是无理数，即它不能表示为一个分数。（我们将在下一章中给出证明。）当时这项发现激起了一片错愕，但时至今日我们已经能够欣然接受：要想将正方形对角线的长度之类的事物模型化，就必须扩充我们的数系。抽象方法再一次使我们的任务变得轻易。我们引入一种新的符号——$\sqrt{2}$，并引入一条说明它能够做什么的新规则：它的平方等于2。

如果你对此颇有研究，你可能会反对我刚才说的，理由是，这样的规则并不能区分$\sqrt{2}$和$-\sqrt{2}$。处理这个问题的办法之一就是向我们的数系中引入一个新的概念——序。比较数和数之间的大小总是有用处的，而且这还能使我们通过$\sqrt{2}$的额外的性质——大于0——来识别它。不过即便没有这种性质，我们也已经能够做一些运算，例如：

$$\frac{1}{\sqrt{2}-1}=\frac{\sqrt{2}+1}{(\sqrt{2}-1)(\sqrt{2}+1)}=\frac{\sqrt{2}+1}{(\sqrt{2})^2-\sqrt{2}+\sqrt{2}-1}$$

$$=\frac{\sqrt{2}+1}{2-1}=\sqrt{2}+1$$

而且，不区分$\sqrt{2}$和$-\sqrt{2}$其实是有一点好处的——上面的计算对这两个数都是成立的。

从用来描述数系每次扩充所得到的新数的名字，我们能够发现在历史上对抽象方法质疑的一些痕迹，比如“负的”和“无理的”。但更让人难以下咽的还在后面，这就是“虚幻的”，或者“复杂的”数，即形如$a+bi$的数，其中a和b均为实数，i是-1的平方根。

从具象的观点来看，我们会很快就摈弃-1的平方根：因为任何数的平方都是正的，-1根本就没有平方根，故事到此为止。然而，若采纳了抽象的观点，这种反对意见就显得软弱无力了。只要引入方程$x^2=-1$的解并把它称作i就好了，为什么不继续单纯地把数系扩充下去呢？为什么偏偏它的引入就应该比之前的$\sqrt{2}$更值得反对呢？

一种回答大概是，$\sqrt{2}$能够按十进制小数展开，（原则上）能够计算到任意精度，而i就与此不同了。但这说的只不过是我们已经知道的事

情，即i不是实数——正如$\sqrt{2}$不是有理数一样。这并不能阻挡我们扩充数系，在其中进行如下的运算：

$$\frac{1}{i-1}=\frac{i+1}{(i-1)(i+1)}=\frac{i+1}{i^2-i+i-1}=\frac{i+1}{-1-1}=-\frac{1}{2}(i+1)$$

i和$\sqrt{2}$之间最主要的区别就是我们被迫抽象地去思考i，而对于$\sqrt{2}$我们则还有备选方案，可以将它具体地表示为1.4142…，或者看作单位正方形的对角线长度。要看出为什么i没有这样的表示方法，不妨问问自己这个问题：-1的两个平方根中，哪个是i哪个是-i呢？这个问题是没有意义的，因为我们对i所定义的唯一的性质就是平方等于-1。既然-i也有同样的性质，那么关于i成立的那些命题，如替换为关于-i的相应命题，必定依然成立。一旦领会了这一点，就很难再赞同i指示一个独立存在的实在的客体。

这和一个著名的哲学难题有相似之处。你对红色所产生的感受与我对绿色产生的感受（交换亦可）有没有可能是相同的呢？一些哲学家很严肃地思考这个问题，并定义“感受性”一词来表示我们所拥有的绝对的内在体验，比如我们对色彩的体验。而另一些人并不相信感受性。在他们看来，“绿色”这样的词有更抽象的定义，那就是根据它在语言系统中所发挥的作用，也就是说，根据它与“草地”、“红色”等概念之间的关系。因此，就这个论题，要想从人们谈论色彩的方式来推断出他们的态度是不可能的，除非在哲学争论当中。类似地，在实践中，关于数和其他数学对象，重要的只是它们所遵循的规则。

如果说为了使方程$x^2=-1$有解我们引入i，那么其他类似的方程呢？比如$x^4=-3$或者$2x^6+3x+17=0$呢？值得注意的是，人们发现，所有这样的方程都可以在复数系中求解。也就是说，我们通过接受i作出小小的投资，结果得到了许多倍的回报。发现这个事实的历史过程有点复杂，但人们通常将它归功于高斯。这个事实被人们称为代数基本定理，它给我们提供了令人折服的证据，使我们相信i的确有合情合理、自然而然的地方。我们的确无法想象一个篮子里有i只苹果，车行途中经过了i个小时，银行账户透支了i英镑。但对数学家来说，复数系已经必不可少。对科学家和工程师同样也是。比如，量子力学的理论就高度依赖于复数。复数作为最佳的例证之一，向我们表明了一条概括性原则：一种抽象的数学构造若是充分自然的，则基本上必能作为模型找到它的用途。

初探无穷大

一旦我们学会抽象地思考，事情就会立刻变得令人愉悦，这个境况有点像突然能够骑自行车而不必去担心保持平衡。然而，我也并不希望使读者觉得抽象方法就好像是印钞许可证。我们可以来作一个有趣的对比，比较一下向数系中引入i与引入数字无穷大之间的区别。乍看起来，似乎没什么可以阻止我们的：无穷大应当用来表示1除以0之类的，所以，为什么不使∞成为一个抽象符号，用它来表示方程0x＝1的解呢？

但当我们想做算术时，这个想法的问题立刻就来了。我们在这里举个例子，利用乘法结合律**M2**和0×2＝0的事实，就可以得到简单推论：

$$1=\infty\times 0=\infty\times(0\times 2)=(\infty\times 0)\times 2=1\times 2=2$$

这个式子表明，方程0x＝1的解若存在将会导致不相容性。这是否意味着无穷大不存在呢？并不是。这只说明，无穷大的自然概念与算术定律是不相容的。扩充数系以将符号∞包含进来，并且接受在新的系统下这些算术定律并非总是成立，这样做有时是有用处的。但是通常，人们还是希望保持算术定律，不考虑无穷大。

把负数和分数放到指数上

抽象方法还有一点极大的优越性：它使我们能够将熟悉的概念扩展到不熟悉的情况下，赋予其新的意义。我所说的“赋予意义”的确是恰当的，因为我们所做的正是去赋予意义，而不是去发现某种早就存在的意义。这当中有一个简单的例子，那就是我们如何扩展指数的概念。

如果n是个正整数，那么a^n即表示n个a相乘的结果。如$5^3=5\times5\times5=125$以及$2^5=2\times2\times2\times2\times2=32$。但若以此作为定义，我们就不容易去解释$2^{3/2}$这样的表达式，因为你不可能拿出一个半的2，把它们乘在一起。那处理这种问题的抽象方法是什么呢？我们又一次地需要抛开寻找内在意义的意识。在本例中即需要忽视a^n的内在意义，转而考虑关于它的规则。

关于指数运算的两条基本规则是：

E1 对任意实数a，$a^1=a$。

E2 对任意实数a和任意一对自然数m、n，有$a^{m+n}=a^m\times a^n$。

例如，$2^5=2^3\times2^2$，因为2^5表示$2\times2\times2\times2\times2$，而$2^3\times2^2$表示$(2\times2\times2)\times(2\times2)$，由乘法结合律可知两数相同。

从上述两条规则出发，我们可以迅速重新得到已经知道的一些事实。比如，根据**E2**即知$a^2=a^{1+1}$等于$a^1\times a^1$；再根据**E1**，此即为$a\times a$，正如我们所了解的。除此以外，我们现在还能够做更多的事情。让我们用x来表示$2^{3/2}$。那么$x\times x=2^{3/2}\times2^{3/2}$，由E2得知它就是$2^{3/2+3/2}=2^3=8$。也就是说$x^2=8$。这并没有完全确定下x，因为8有两个平方根。所以我们通常会采取如下的准则。

E3 如果$a>0$且b是实数，那么a^b为正数。

再应用上**E3**，我们就发现$2^{3/2}$是8的正平方根。

这并不是对$2^{3/2}$的“真正值”的发现。但这也不是我们对表达式$2^{3/2}$的随意解读——如果我们希望保持规则**E1**、**E2**和**E3**，这就是唯一的可能性。

用类似的办法，我们可以对a^0给出解释——至少当a不为0时。由**E1**和**E2**，我们知道$a=a^1=a^{1+0}=a^1\times a^0=a\times a^0$。消去律**M5**指出，无论a取何值，都有$a^0=1$。对于负指数，如果我们已经知道了$a^b$的值，那么$1=a^0=a^{b+(-b)}=a^b\times a^{-b}$，由此推出$a^{-b}=1/a^b$。例如，$2^{-3/2}$就等于$1/\sqrt{8}$。

对数是另一个抽象地来看会变得更加容易的概念。关于对数，我在本书中要说的不多。但如果它确实困扰你，那么你可以消除顾虑，只要了解它们遵循如下三条规则就足以使你去应用对数了。（如果你希望对数是以e为底而不是以10为底的，只需要在**L1**中把10替换为e即可。）

L1 log（10）＝1。

L2 log（xy）＝log（x）＋log（y）。

L3 若x＜y，则log（x）＜log（y）。

例如，要得到log（30）小于$3/2$，可以应用**L1**和**L2**得出log（1000）＝log（10）＋log（100）＝log（10）＋log（10）＋log（10）＝3。而由**L2**得出2log（30）＝log（30）＋log（30）＝log（900），又由**L3**得到log（900）＜log（1000）。因此2log（30）＜3，即得log（30）＜32。

在这本书的后面，我还将讨论许多类似性质的概念。试图具体地理解它们会让你感到困惑，但当你放轻松些，不再担心它们是什么并且应用抽象的方法，那这些概念的神秘性就消失了。

第三章
证明

图9中画出了五个圆，第一个圆上标出了一个点，第二个圆上标出了两个点，依此类推。连结这些圆上点的所有可能直线段都画了出来，这些线段将圆分割成若干区域。去数一数每个圆的区域数的话，会得到序列1，2，4，8，16。我们可以立刻辨别出这个序列：似乎圆上每添加一个新的点，区域个数就会加倍，因而n个点就分出了2^{n-1}块区域——至少在没有三线共点的情况下。

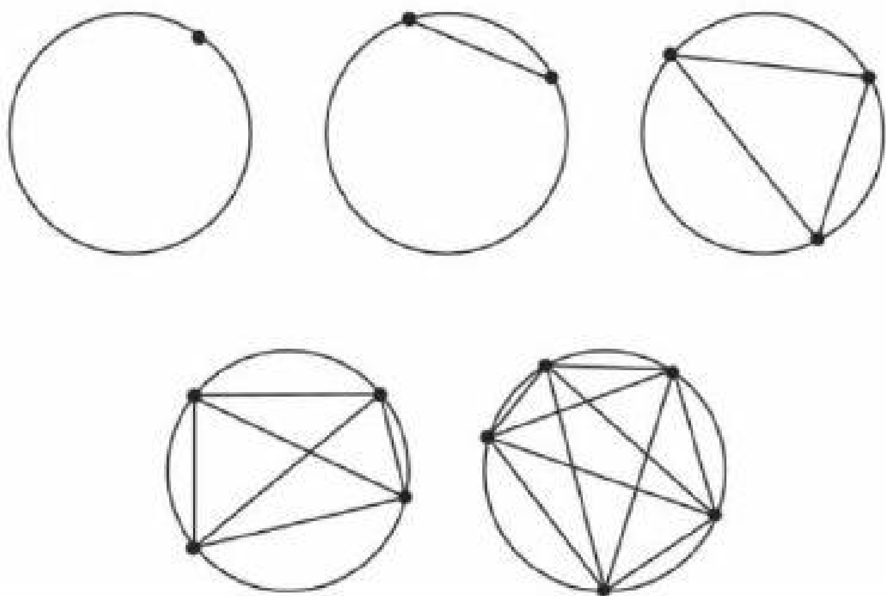


图9 将圆分割成若干区域

然而，数学家很少会对“似乎”这样的用语感到满意。他们所需要的是证明，也就是能够扫清一条论断中所有疑点的论证。可是，这究竟是什么意思呢？尽管我们常常可以立论，在虑及所有情理之中的怀疑时认为论断正确；但如果要下结论说我们的论证扫清了一切的疑点，那我们必定要更上一层楼。历史学家能够给出许多例子来说明，一些论断一度被认为毋庸置疑，而后来却被证明是错误的，其中有一部分就是数学方面的论断。为什么当今数学中的定理与此会有所不同呢？我在下面将回答这一问题，给出几个证明的例子来，并从中概括出一些一般性的结论来。

根号2的无理性

我在上一章中说到，一个数如果可以写成分数p/q（其中p和q是整数）的形式则称为有理数，若不可以则称为无理数。数学中一条著名的证明表明了 $\sqrt{2}$ 是无理数。这项证明阐释了反证法这种技术，即通过推出矛盾来证明。

这样的证明以假设要证的结论为假开始。这看似有点奇怪，但其实我们在日常对话中经常用到这一技巧。如果你去警局报案，声称目睹了一辆车被人故意破坏，别人指控你自己就是破坏者，那你很可能说："如果是我干的，那我根本不可能以这种方式让自己成为目标。"你暂且采纳了你就是破坏者的（非真）假设，以此来说明这是多么荒唐。

我们要去证明 $\sqrt{2}$ 是无理数，那就先假设它是有理数，再来表明这一假设会引出荒诞的结论。我在下面将按步骤写出证明，给出比较多的细节，很多读者也许并不需要这么多。

1. 如果 $\sqrt{2}$ 是有理数，那么我们可以找到整数p和q使得 $\sqrt{2}$ ＝p/q（由"有理数"的定义）。
2. 任意分数都能够写成某个分数r/s，r和s不全是偶数。（分子分母连续除以2，直到至少有其中一个变成奇数。例如分数1412/1000等于706/500等于353/250。）
3. 因此，如果 $\sqrt{2}$ 是有理数，我们就可以找到两个不全为偶数的整数r和s使得 $\sqrt{2}$ ＝r/s。
4. 如果 $\sqrt{2}$ ＝r/s，则2＝r2/s2（等式两端平方）。
5. 如果2＝r2/s2，则2s2＝r2（等式两端乘以s2）。
6. 如果2s2＝r2，则r2是偶数，即r必须是偶数。
7. 如果r是偶数，那么存在某个整数t使得r＝2t（由"偶数"的定义）。
8. 如果2s2＝r2且r＝2t，则2s2＝（2t）2＝4t2，于是得到s2＝2t2（两端除以2）。
9. 如果s2＝2t2，那么s2是偶数，意味着s是偶数。
10. 按照 $\sqrt{2}$ 是有理数的假设，我们已经表明 $\sqrt{2}$ ＝r/s，r和s不全是偶数（第3步）。我们之后又得到r是偶数（第6步），s是偶数（第9步）。这是一个明显的矛盾。因为 $\sqrt{2}$ 是有理数的假设会推出明显错误的结论，所以这个假设本身必定是错误的。因此 $\sqrt{2}$ 是无理数。

以上我尽可能使每一步推导都做到明显有理有据，从而使结论无可反驳。但是，我真的完全没有给怀疑留下余地吗？若有人愿意跟你打赌，如果找不到两个整数p和q使得 $p^2＝2q^2$ 就给你一万英镑，但如果找到了就处死你，那你愿意接受挑战吗？如果你愿意，又是否会有一点点的不安呢？

第6步中包含了 r^2 是偶数则r必定是偶数的论断。这看起来相当明显（奇数乘以奇数是奇数），但如果我们要想使 $\sqrt{2}$ 是无理数这一命题绝对确定无疑，或许可以再加强一些。让我们把它分解成五个子步骤：

6a. r是整数，r2是偶数。我们要表明r必定也是偶数。让我们假设r是奇数，然后寻求矛盾。

6b. 因为r是奇数，所以存在整数t使得r＝2t＋1。

6c. 于是推出r2＝（2t＋1）2＝4t2＋4t＋1。

6d. 但是4t2＋4t＋1＝2（2t2＋2t）＋1，这是奇数，与r2是偶数的事实矛盾。

6e. 因此r是偶数。

现在步骤6完全滴水不漏了吗？可能还没有，因为子步骤6b仍需要证明。毕竟，奇数的定义仅仅是非2的倍数的整数。为什么每个整数要么就是2的倍数，要么就比2的倍数多1呢？我们可以用以下论据来表明它。

6b1. 如果某个整数r是2的倍数，或者比2的倍数多1，我们就称它为一个好数。如果r是好数，则r＝2s或r＝2s＋1，其中s也是整数。如果r＝2s，则r＋1＝2s＋1；如果r＝2s＋1，则r＋1＝2s＋2＝2（s＋1）。不管是两种情况中的哪一种，都有r＋1也是好数。

6b2. 1是好数，因为0＝0×2是2的倍数，且1＝0＋1。

6b3. 重复利用6b1这一步，我们可以推出2是好数，然后3是好数，然后4是好数，等等等等。

6b4. 因此，所有正整数都是好数，这正是我们要证明的结论。

我们现在完成了吗？大概这一回最不牢靠的步骤要数6b4了，因为它是从前一步“等等等等”这样十分模糊的字眼中得到的。步骤6b3告诉我们怎样去表明任意给定的正整数n是好数。但麻烦在于，若按照上面的论证，我们需要从1一直数到n，如果n很大的话，这就要花很长时间。如果我们想要说明所有正整数都是好数，情况就更糟糕了，看起来这样的论证似乎永无结束之日。

但另一方面，鉴于步骤6b1到6b3实实在在、确切无疑地给我们提供了一种方法，能够说明任意的n都是好数（只要我们有时间），这一反对意见看起来就不合理。实际上它是如此不合理，以至于数学家们采纳了下述原则作为一条公理。

假设关于任意正整数n有一陈述s（n）。（在我们的例子中，s（n）即表示陈述“n是好数”。）如果s（1）为真，而且s（n）为真总蕴含s（n＋1）为真，那么s（n）对任意n都为真。

这就是数学归纳法原理，熟悉它的人也简称它为归纳法。通俗地讲，它所说的其实就是，如果你有一列无穷多的陈述序列想要证明，那有一种办法：证明第一条为真，并且每一条都蕴含下一条。

上述几段内容说明了，数学论证中的每一步都可以分解成更小的，因而也更加清晰有据的子步骤。这些小步骤又可以进一步分解为子子步骤，等等。数学中有个根本性的重要事实，那就是这样的过程最终必然会终止。原则上，如果不断地将步骤分解为更小的步骤，你最终会得到一条非常长的论证，它以普遍接受的公理开始，仅通过最基本的逻辑原则（例如“若A为真且A蕴含B，则B为真”）一步步推进，最终得到想要求证的结论。

上一段中我所说的远非显然：事实上，这正是20世纪早期的重要发

现，很大程度上归功于弗雷格、罗素和怀特海（参见“延伸阅读”）。这一发现对数学产生了深远的影响，因为它意味着，任何关于数学证明有效性的争论总是能够解决的。而在19世纪，与此形成对比的是，的确存在着关于数学实体问题的真正分歧。譬如，现代集合论之父格奥尔格·康托尔，基于某个无穷集可能比另一个无穷集“更大”这样的思想提出了一些论点。今天人们已经接受了这些论点，但当时的人们却对此产生强烈的怀疑。如今，如果人们对于某个证明的正确性存在分歧，那要么是因为这个证明写得不够详细，要么是因为人们还没有付出足够努力来仔细地理解、检查这个证明。

不过这并不意味着分歧永远不会出现。比如一种常见的情况：某人炮制了一篇极长的证明，在某些地方不够清楚，同时还包含许多小错误，但并非明显可见的根本性错误。要想下结论说这样的论点是否能做到滴水不漏，通常要耗费巨大的工作量，而这样的工作却没有太高的回报。即使是提出证明的人自己可能也不愿意做，生怕会发现他的证明是错的。

虽然如此，争论在原则上必然能够解决这一事实的确使数学独一无二。没有任何一个学科像数学一样具有这一特性：有些天文学家仍然固守着宇宙的稳态理论；关于自然选择究竟有多大的解释力，生物学家各自都抱有极为不同的坚定信念；关于意识与物质世界的关系，哲学家们具有根本性的分歧；经济学家也追随着观点截然相反的不同学派，如货币主义和新凯恩斯主义。

理解前面“在原则上”这个短语是很重要的。没有哪个数学家愿意费心写出证明的完整细节——从基本公理开始，仅通过最明显、最易于检查的步骤来推导出结果。即使可行，也并不必要：数学论文是写给经过严格训练的读者的，无须事无巨细详细说明。而如果某人提出一个重要的论断，其他数学家发觉难以理解其证明，他们就会要求作者详细解释。这时，把证明步骤分解为较易理解的小步骤的过程就会开始。同样因为听众都是经过严格训练的，这个过程通常不需要进行太久，只要给出了必要的解释或者发现了其中的错误就可以了。因此，对某个结果的一个证明如果的确是正确的，那几乎总会被数学家当作是正确的。

一部分读者可能会萌生这样的问题，我还未触及：为什么我们应该接受数学家提出的公理呢？比方说，如果有人反对数学归纳法原理，我们应当怎样回应呢？大多数数学家会给出如下的答复。首先，所有理解了归纳法的人应该都认为它是显然合理的。其次，公理系统的主要问题并不是公理的真实性，而是公理的自洽性和有用性。数学证明实际上所做的正是要表明，由特定前提——如数学归纳法，能够得到特定的结论——如$\sqrt{2}$是无理数。这些前提假设是否正确则是与此完全无关的问题，我们可以安然地把它们留给哲学家。

黄金分割比的无理性

人们在学习高等数学时，走到一个证明的结尾处，通常会经历这样的思考："我理解每一行是怎样由前一行得到的，但是我却不明白为什么这个定理是正确的，人们是怎样想到这个论证的。"我们经常想从证明中得到更多的东西，而不仅仅是确信它的正确性。读过一个好的证明之后，我们会感到它对定理进行了一番阐明，使我们理解了之前所不理解的一些东西。

由于人类大脑有很大一部分是用于处理视觉数据的，我们也就不难理解有很多论证都用到了我们的视觉能力。为了说明这一点，我将给出另一个关于无理性的证明，这一回是所谓的黄金分割比。几百年来，这个数一直使非数学家为之着迷（数学家也着迷，只是程度较轻）。它是具有如下特征的矩形的长宽比：

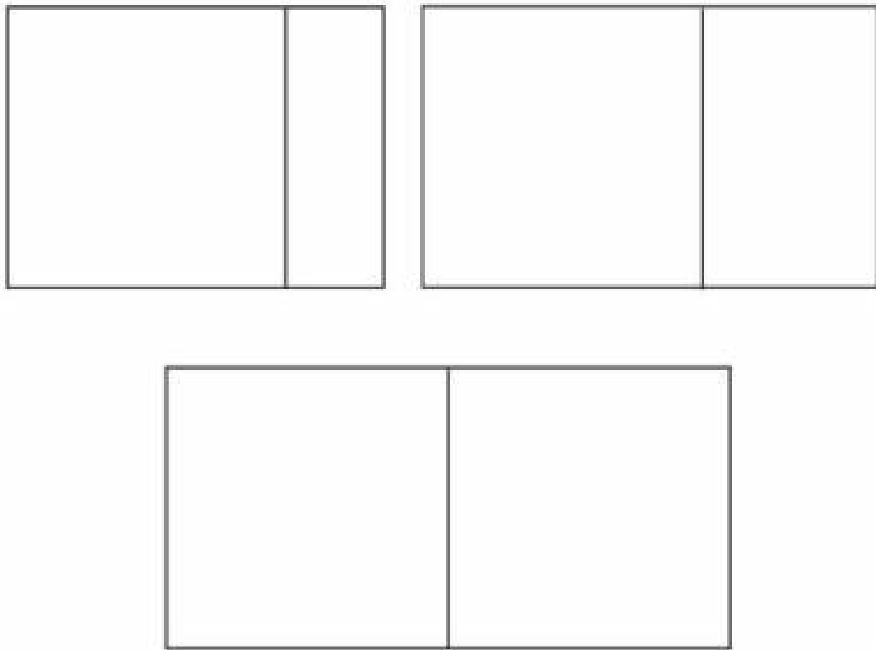


图**10** 黄金分割的存在性

从矩形中切掉一个正方形，剩下一个小矩形，它旋转之后恰与原始矩形的形状完全一样。图10中的第二个矩形正是这种情况。

为什么这样的边长比是存在的？（数学家已经被训练得惯于提出这类问题。）要回答它可以换一个角度，想象一个小矩形从正方形的一侧生长出来，使整个图形变成一个大矩形。刚开始的时候，小矩形很长很瘦，而大矩形仍然差不多是个正方形。如果我们让小矩形继续生长，直到它自己变成一个正方形，那大矩形的长就是宽的两倍。所以，小矩形最开始的时候比大矩形瘦得多，现在却比大矩形胖（相对于各自的尺寸来讲）。在这两者之间必存在一点，使得两个矩形形状相同。图10说明了这样的过程。

还可以用另一种方式来考虑黄金分割比的存在，那就是把它算出来。如果我们把它叫作x，并假设正方形的边长为1，那么大矩形的边长

就是1和x，小矩形的边长是x-1和1。如果它们的形状相同，那么

$$x=\frac{x}{1}=\frac{1}{x-1}$$

。两端同时乘以x-1，我们就得到x（x-1）＝1，所以x^2-x-1＝0。求解这个二次方程，并且知道x不是负数，我们就得出

$$x=\frac{1+\sqrt{5}}{2}$$

。（如果你在数学上很有造诣，或者深入理解了上一章，你可能会问为什么我如此确信5姨存在。实际上，这第二种论证正是要把几何问题归结于一个等价的代数问题。）

得出了比例x存在之后，让我们来考虑对边长为x和1的矩形进行如下的操作。首先，从中切掉一个正方形，由黄金分割比的定义，剩下的小矩形与原来的矩形形状相同。然后再不断地重复这一基本操作，得到一系列越来越小的矩形，每一个的形状都与之前的形状相同，因而长宽比都是黄金分割比。很显然，这一过程永远不会终止。（参见图11的第一个矩形。）

现在让我们对长宽比为p/q的矩形施行同样的操作，其中p和q都是整数。也就是说，这个矩形与边长为p和q的矩形形状相同，因而可以被分成p×q个小正方形，如图11中的第二个矩形所示。如果我们从这个矩形的一端切掉一些大正方形会怎么样呢？如果q小于p，那么我们切掉一个q×q的正方形，得到一个q×（p-q）的矩形。我们可以继续切掉下一个正方形，依此类推。这个过程会永远持续下去吗？不会。每次切掉的正方形都是小方格的整数倍，我们不可能操作多于p×q的次数——因为一开始就只有p×q个小方格。

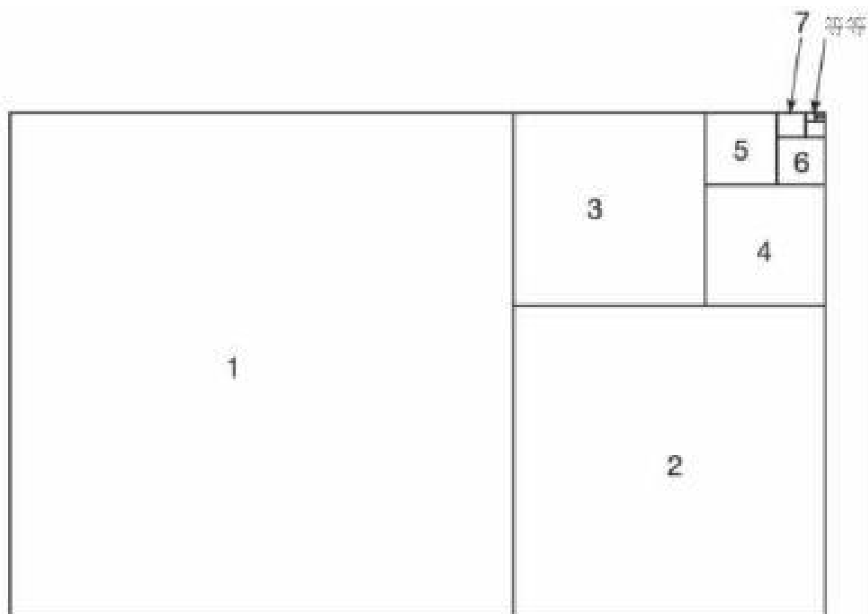
我们表明了如下两个事实。

1. 如果矩形的边长比是黄金分割比，那么我们可以一直从中切掉正方形，没有尽头。

2. 如果矩形的边长比是某对整数p和q的比值p/q，那么 就不能永不停止地从中切掉正方形。

于是我们可以得到，比值p/q不是黄金分割比，无论p和q取什么值。换言之，黄金分割比是无理数。

仔细思考上面这个证明，你最终会发觉，虽然乍看它和之前对$\sqrt{2}$是无理数的证明很不一样，但实际上也并无太大区别。不过，我们将它呈现出来的方式确实不同——而且对很多人来说，这种方式更有吸引力。



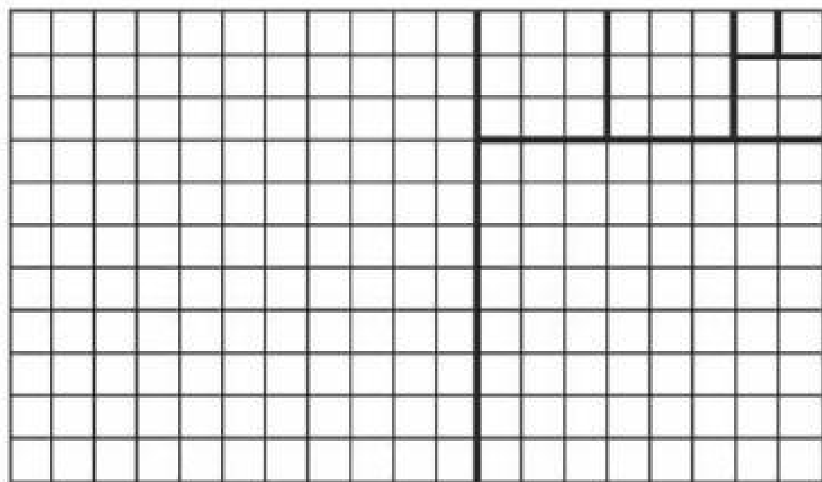


图11 从矩形中切掉正方形

圆的分割

既然我已经谈到了数学证明的本质，现在让我们回到本章开始时的问题。圆上有n个点，我们用直线将这些点两两连结起来，希望能够表明这些直线所分割出的区域总数是2^{n-1}。对于n为1，2，3，4或5的情形，我们已经看出这是对的。为了一般性地证明这个陈述，我们很想找到一种令人信服的推理，能够说明边界上每增加一点区域总数就增加一倍。这样的推理会是什么样的呢？

脑子里无法立刻蹦出些什么来，那么我们可以通过观察被分割的圆的图形来着手，看看能否从中发现某种模式并提炼概括。例如，圆上有三个点，就产生了三个外围区域和一个中心区域。圆上有四个点，就有四个外围区域和四个中心区域。圆上有五个点，就有一个中心五边形，五个三角形从中心指向外围，五个三角形嵌入这个五角星之中，于是又形成一个五边形，最后还有五个外围区域。因此，我们可以把4看作3＋1，把8看作4＋4，把16看作5＋5＋5＋1。

这能起到什么作用吗？我们似乎还没能掌握足够多的例子，看出清楚的模式，所以让我们再来画出圆上有六个点的情形。图12显示了这种情况。这一回外围有六块区域。每一块都与一个指向中心的三角形相邻。这样的区域两两之间，各夹着两个小三角形区域，至此共有6＋6＋12＝24块区域，还需要继续数位于中心的六边形内包含的区域。里面分成了三个五边形，三个四边形，还有一个中心三角形。所以看起来很自然地，区域总数是6＋6＋12＋3＋3＋1。

可是好像出了点什么错，因为结果是31。我们哪里疏忽了吗？事实上并没有：正确的序列开头几项是1，2，4，8，16，31，57，99，163。其实只要稍加深入思考，我们就能看出，区域总数不可能每次都翻倍。在刚开始时就有点麻烦，圆上0个点得到的区域总数是1而不是12，所以加上第一个点的时候区域总数就不是翻倍。尽管这种例外情形时常发生在0上，大多数数学家还是会认为这是个麻烦。但是，当n是个比较大的数时问题会更加严重，这时2^{n-1}显然会特别大。比如当n＝20时2^{n-1}是524 288，当n＝30时，结果是536 870 912。在圆的边上画30个点就会把圆分割成超过5亿个不同区域，这可能么？当然不可能。想象一下，在一块地上画一个大圆，在圆上打30个间隔不一的桩子，用很细的线把桩子两两相连。结果得到的区域数量确实比较大，但也不会大到难以想象的程度。如果圆的直径是10米，把它分成5亿块，平均每平方厘米里就会有超过600块区域。这个圆一定密密麻麻布满了线，但周围只有30个点，实际情况显然不是这样的。

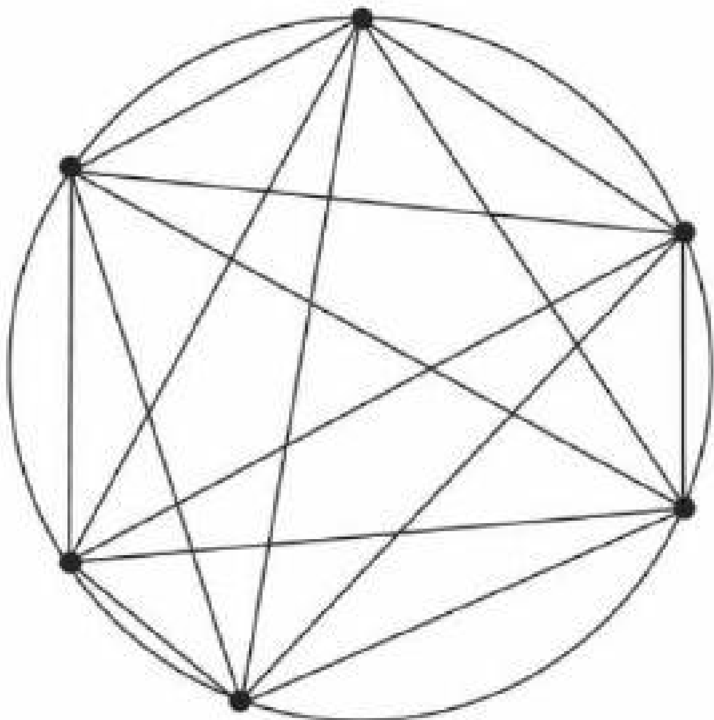


图12 圆的分割

我前面说过，数学家对待“显然”这样的词非常慎重。但在这个例子中，我们的直觉能够以坚实的论证来支持，归结如下。如果圆被分割成数量巨大的多边形区域，这些区域之间必然有大量的顶角。每个顶角处都是两条线的交点，和这个交点相联系的桩子，也就是两条线在圆上的端点有4个。我们在圆上选取这样的4个桩子，第一个桩子有30种可能的选择，第二个有29种，第三个有28种，第四个有27种。这意味着，选取4个桩子的方式共有$30\times29\times28\times27=657\ 720$种，但是这样就没有考虑到，如果以不同的顺序选取了同一组桩子，就重复列入了相同的交点。选出同一组4个桩子共有$4\times3\times2\times1=24$种不同方式，考虑到这一点，我们就能够算出交点总数是657 720/24＝27 405个，根本不可能像536 870 912块区域的顶角数量那样巨大。（实际上，30个点分割出的区域的真正数量是27 841。）

在这个让我们引以为戒的故事中，包含了很多与证明数学陈述相关的重要教训。最明显的一个就是，如果不去小心地证明你所说的话，那你就有说错的危险。一个更积极一点的寓意是，如果确实努力去证明一个陈述，那你将能以全然不同而且更有意思的方式理解它。

毕达哥拉斯定理

毕达哥拉斯的著名定理所讲的是，假设一直角三角形的三边长为a、b和c，其中c是斜边长（直角所对的边），则$a^2+b^2=c^2$。这个定理有若干种证明，其中有一种特别简短，而且很容易理解。它差不多只需要下面两幅图即可。

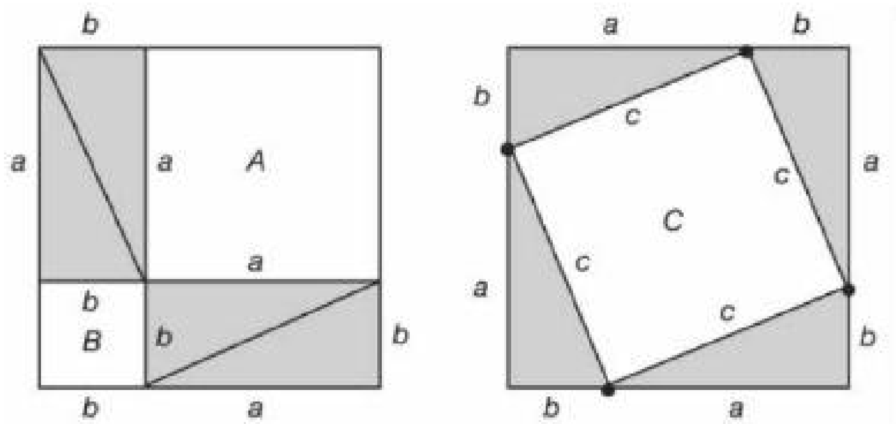


图13 毕达哥拉斯定理的简短证明

图13中，我标为A、B、C的正方形，边长分别为a、b、c，因而面积分别为a^2、b^2、c^2。因为四个三角形的移动并没有引发面积的改变，也没有使它们重合，所以两幅图中，大正方形去掉四个小三角形所得的面积应当是相同的。但在左图中，这个面积是a^2+b^2，而右图中则是c^2。

缺角正方形网格的铺地砖问题

这里有个著名的难题。画八横八纵正方形网格，去掉相对的两个角。你能用多米诺骨牌形状的地砖——每一块正好覆盖两个相邻方格，把剩余部分覆盖吗？我在图中（图14）表明，如果用四横四纵来代替八横八纵，你是办不到的。假设你决定用一块地砖覆盖我图中标为A的区域，那么容易看出，你必须还要把地砖放到B、C、D和E的位置上，剩下一个小方格无法覆盖。既然右上角的格子无论如何总要覆盖住，而仅剩的另一种覆盖的方式也会导致类似的问题（通过位置的对称关系），所以覆盖整个图形是不可能的。

如果我们用五来代替四，网格的铺设仍然是不可能的，原因很简单，每块地砖占两个方格，而小方格总数是23——是个奇数。但是8^2-2＝62是个偶数，所以我们不能把这样的论证用于八横八纵的方格。另一方面，若想找到一种类似于刚才四横四纵情况中我给的证明，你很快就会放弃，因为你所需要考虑的可能情况实在太多了。那对这个问题应该怎样入手呢？如果你从没接触过这个问题，我强烈建议你在继续阅读之前先尝试求解一下，或者暂时跳过下一段，因为如果你解决了它，你将能够对数学中的愉悦感有很好的理解。

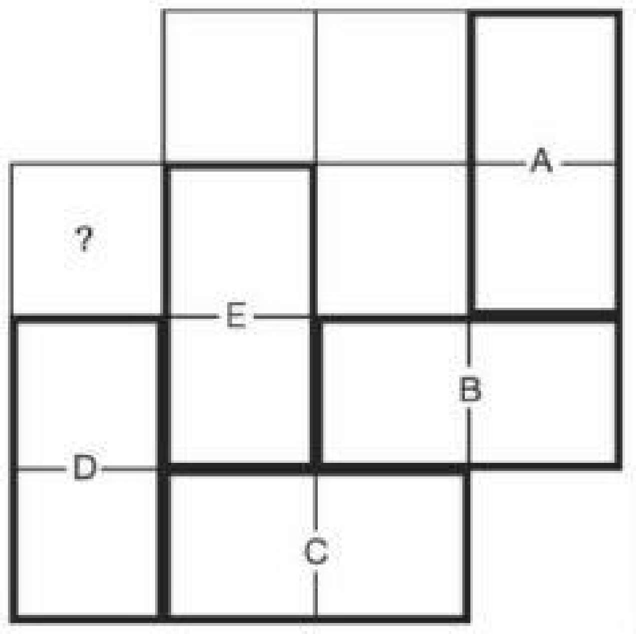


图14 缺角正方形网格的铺地砖问题

对于无视我建议的读者——经验表明这些人占大多数，有一个词几乎可以代表全部证明：国际象棋。国际象棋的棋盘是八横八纵的方格。每个小格交错地填上黑色和白色（就象棋游戏本身来考虑似乎并不是必要的，不过能够使视觉看起来更轻松）。两个相对的顶角方格颜色是相同的。如果它们都是黑色的，那么一旦把它们去掉，剩下的棋盘就有32个白格子和30个黑格子。每一块骨牌只能覆盖两种颜色的方格各一块，所以一旦你放进了30块骨牌，无论是怎么放的，最终都必然剩下两个白方格，它们是无法覆盖的。

这个简短的论证极好地表明，证明何以能够不仅仅保证陈述的正确性。例如，四乘四方格去掉相对两角无法被覆盖，这条陈述我们现在有了两种证明。一种是我前面给出的，另一种是四乘四版本的象棋盘论证法。这两种证明都得到了我们想要的结果，但只有第二种给了我们一种关于无法覆盖的类似推理的东西。这样的推理能够立即告诉我们，一万乘一万的方格去掉相对的两角也是无法覆盖的。第一种论证则只能告诉我们四乘四的情况。

第二种论证有个值得瞩目的特征，它完全依赖于一种思想，这种思想虽出人意料，但一经理解便显得非常自然。人们经常很困惑，为什么数学家有时会用“优美”、“漂亮”甚至“绝妙”来形容一些证明。这样的例子就让我们对其含义有了一点理解。音乐也能够提供一个有用的类比：一段乐曲刚开始可能沿意想不到的和声方向行进，过后却感觉到非常完美恰当，或者一段管弦织体呈现出整体大于部分之和的境界，其方式我们还无法全然理解——每当这些时候我们就会为之陶醉。在数学证明中，有突如其来的启发，有出人意料却自然而然的思想，还有引人入胜、有待进一步发掘的暗示，这些都能够给我们带来类似的愉悦感。当然，数学的美不同于音乐的美，可音乐的美同样也不同于绘画的美、诗歌的美、姣好面容的美。

三条看似显然实则需要证明的陈述

较高等的数学中，有一点让很多人感到费解：其中有一些定理看上去非常显然，简直无须证明。遇到这样的定理时，人们常常会问："如果这都不算显然，那还有什么才算呢？"我一位先前的同事对此给了一个很好的回答：如果脑子里立刻就有证明，那么这条陈述才是显然的。在本章的剩余部分，我将给出三条陈述作为例子，它们看上去都是显然的，却无法通过这样的检验。

1. 算术基本定理的内容是，每个自然数有且仅有一种被写为素数乘积的方式，不考虑先后顺序。例如，$36=2\times2\times3\times3$，$74=2\times37$，再如101本身就是一个素数（在这样的语境下，它就是单个素数的"乘积"）。观察过这些较小自然数，人们马上就会确信，根本不可能有两种不同方式把一个数表示为素数乘积。这就是这个定理看似几乎无须证明的主要论点。

但它真的有那么明显吗？7，13，19，37，47都是素数，那么，如果算术基本定理是显然的，$7\times13\times19$不等于37×47也应该是显然的。我们当然可以检验出，这两组数确实不同（任何数学家都会告诉你，其中一个比另一个更有意思[1]），但这并不说明它们显然会不相等，也不能解释为什么我们不能另外找到两种素数乘积得到相同的结果。实际上，这个定理并没有很简单的证明方法。如果你脑子里立刻就有了证明，那你的脑子一定很不寻常。

2. 设想你用一段正常的绳子打了一个宽松的结，然后把绳子两端熔合在一起，得到了如图15所示的形状即数学家所称的三叶结。在不把绳子切断的情况下，有没有可能把这个结解开呢？当然不可能。

但为什么我们倾向于说"当然"呢？我们有没有立刻就想到一种论证？可能有——看起来，任何解开结的努力似乎都难以避免地会使得这个结更为繁复。然而，要把这样的直觉转换为有效的证明却是很难的。所谓显然，只不过是没有简单的方法解开这个结。难点在于需要排除所有的可能性，说明先把结变复杂再最终解开它也是不可能的。应当承认，这看起来不太可能，但是，数学中的确存在着这样的现象，甚至在日常生活中都是存在的：比如，为了把房间收拾干净，常常在开始时有必要先把它弄得更乱一些，而不是把所有东西都塞进橱里。

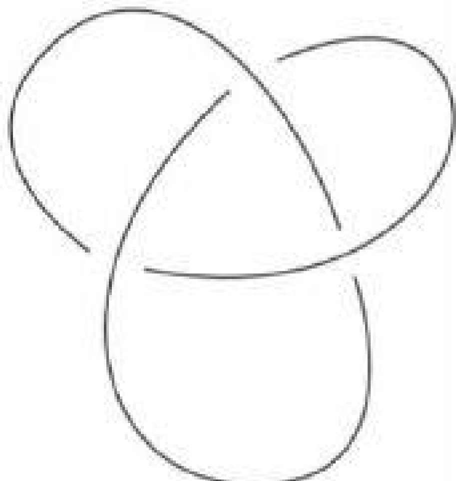


图15 三叶结

3. 平面中的一条曲线，是指笔尖不离开纸面的情况下，你所能画出的任何东西。如果曲线永远不和自己相交，就称它为简单的，如果它最终回到起点处，就称它为闭合的。图16显示了这些定义在实际图形中的反映。图中第一条曲线既是简单的也是闭合的，在平面上围出一块区域，即曲线的内部。很明显，所有简单闭合曲线都将平面分成两部分：曲线里面和曲线外面（如果把曲线本身也看作一部分，就是三个部分）。

这真的有那么明显吗？当曲线不太复杂的时候的确如此。但如图17所示的曲线呢？如果你在靠近图形中心处选一点，点在曲线里面还是外面并不显然。你可能想，大概的确不太显然，可它终归必定会有一个里面和一个外面，哪怕曲线很复杂，在视觉上很难辨别。

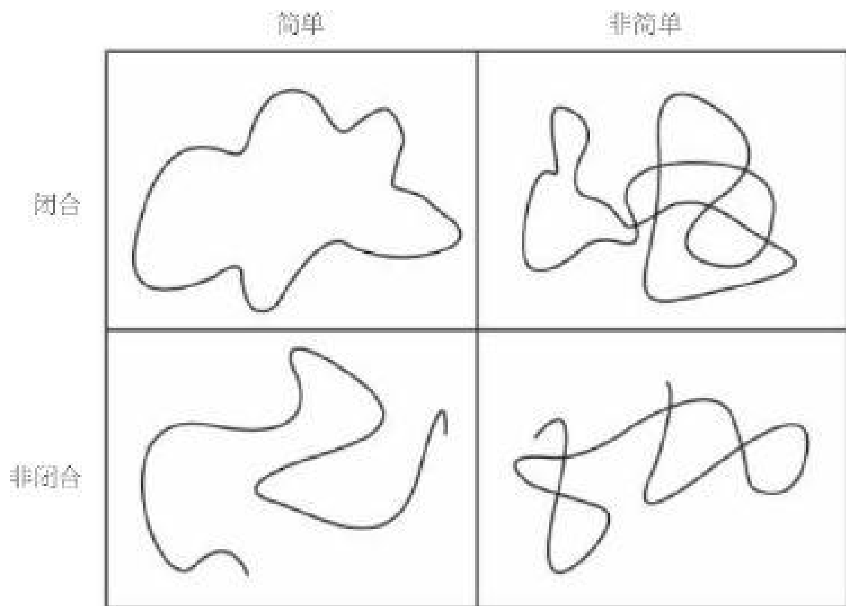


图16 四种曲线

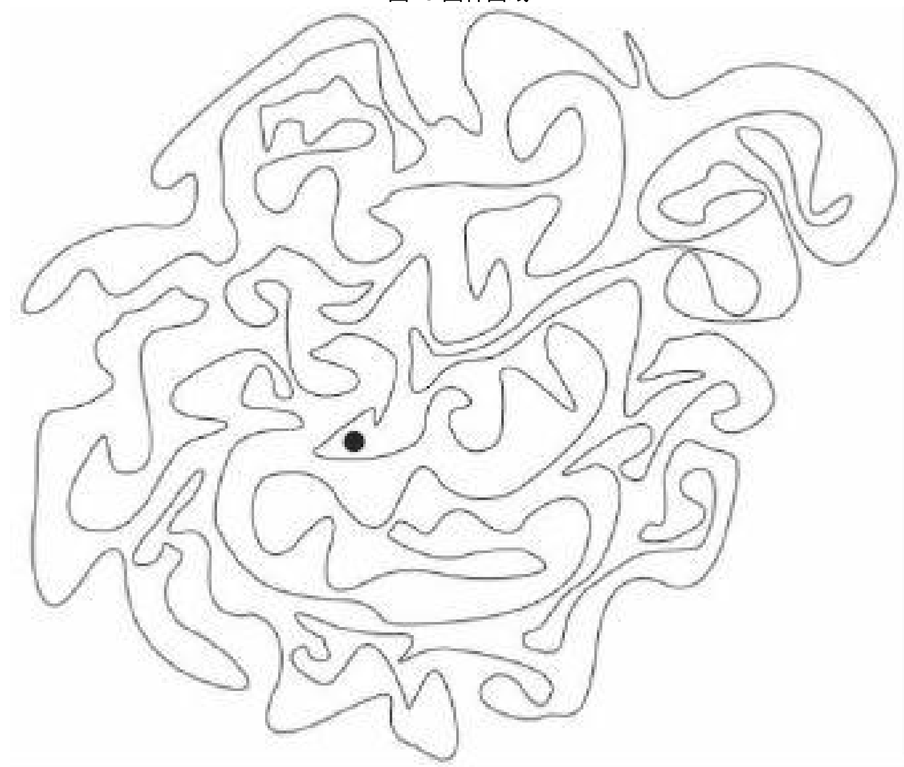


图17 黑点在曲线里面还是外面？

应该怎样来确证这样的看法呢？我们可能会尝试用如下的办法来区别里面和外面。我们暂时假设，确实存在曲线的里面和外面，那么每次穿过曲线的时候，必定要么是从里面穿到外面，要么是从外面穿到里面。于是，如果你想确定点P是在里面还是在外面，你只需从**P**出发画一条线，一直画到远离曲线的另一点**Q**，也就是使**Q**明显位于曲线外面。如果画的这条线和原曲线相交奇数次，那么**P**位于曲线里面，否则位于曲线外面。

这种论述的问题在于，它把许多事情当作理所当然的了。例如，如果你从**P**出发画另一条线，画到曲线外另一点**R**，你怎么知道结果不会不同呢？（结果不会不同，但这需要证明。）实际上，任意简单闭合曲线都有里面和外面这一命题，正是一个著名的数学定理，称作若尔当曲线定理。无论它看起来多么显然，它都需要证明，而且任何我们已知的证明都非常困难，远远超出了本书的论述范围。

【注释】

[1] 英国数学家哈代回忆印度数学家拉马努金时谈到，某次他去医院看望拉马努金，提到路上搭乘的出租车的车牌号为1729，并无特殊之处。拉马努金说："你错了。实际上，这是能以两种方式表示为两数立方和的最小自然数。"

第四章
极限与无穷

在上一章中，我努力指出：原则上，数学证明如何可以完全形式化。如果从特定的公理开始，遵循特定的规则，最后以有趣的数学陈述结束，那么这样的陈述就可以当作定理接受，否则就不能被视为定理。这种思想，即从少数几条公理出发演绎推导出许多复杂的定理，可以追溯到欧几里得。欧几里得只用了五条公理就建立起几何学的主要体系。（关于他的公理，我们将在第六章中讨论。）有人可能提出这样的问题：为什么直到20世纪，人们才认识到这样的思想可以应用于整个数学系统当中呢？

主要的原因可以被归结为一个词：无穷。出于种种原因，无穷这一概念在数学中必不可少，但却很难严格化。在本章中，我将讨论三条陈述。其中每一条乍看起来都普普通通，但经过仔细的考察，会发现最终都涉及到无穷。随之就产生了困难，本章的主要内容就是如何处理这样的困难。

1.2的平方根约为1.414 213 56

上面这条简单陈述仅仅在说一个不大的数差不多等于另一个数，哪里涉及到无穷了呢？答案就藏在“2的平方根”这个短语里面，这个短语隐含假设了2的确存在一个平方根。要想透彻地理解这条陈述，我们就必须问一问，2的平方根究竟是个什么样的对象。于是无穷就来了：2的平方根是一个无穷小数。

注意，下面这条紧密相关的陈述就不涉及无穷的问题：1.414 213 56的平方约为2。这条陈述完全是有限的，但是看起来所谈论的基本上是同样的事情。我们稍后就能看到，这一点至关重要。

说一个无穷小数的平方等于2，这是什么意思呢？在学校里，我们学过有限小数如何相乘，但从没学过无穷小数的乘法——不知出于什么缘故，我们仅仅假设它们能够参与加法和乘法，而不去深究。这运算应该如何完成呢？为了看看会出现何种困难，我们先来考虑加法。当我们把两个有限小数相加时，比方说2.3859加3.1405，我们将一个数写在另一个数下面，从右向左将对应的数位相加。我们从最末端的数位开始做加法——9加5，得到14，于是写下4，并进1。接下来，我们再加倒数第二位——5加0，以及进上来的1，得到6。依此类推，我们就得到了最终的结果5.5264。

现在假设有两个无穷小数。我们不可能从最右端开始，因为无穷小数根本没有最后一位。那么我们要如何把它们相加呢？有一个明显的回答：从左端开始。然而这么做是有缺陷的。譬如，我们再来加有限小数2.3859和3.1405，先加2和3，得到5。接下来再加3和1，得到4，很不幸，这是不对的。

这个错误给我们造成了一点麻烦，不过只要我们保持勇气继续下去倒也没什么大不了的。接下来要相加的两个数字是8和4，我们可以在这第三位上记下一个2，然后将上一位的4改为5加以修正。这个过程还会出现，在第四位我们会先写下一个5，然后修正为6。

注意，这样的修正有可能会在写下该位结果的很久之后才出现。例如，我们要计算1.355 555 555 555 555 557 3加2.544 444 444 444 444 445 2，会首先写下3.899 999 999 999 999 99，但再往下算一位，即7加5时这一连串的9就不得不修正。就像一连串多米诺骨牌一样，这些9一位接一位地变为0。不过这样的计算方法仍然是行得通的，最后得到计算结果3.900 000 000 000 000 002 5，它让我们能够对如何将无穷小数相加赋予意义。不难看出，没有哪一位数需要修正一次以上，所以如果我们有两个无穷小数，那么两数之和的第53位（打个比方），要么是我们在上述过程的第53步写下的那个数，要么是之后修正的那个数——如果有必要修正的话。

我们想要对这一论断赋予意义：存在一个无穷小数，其平方为2。要做到这一点，我们必须首先看一看这个无穷小数是怎样产生的，然后再去理解让它和自己相乘意味着什么。可以预料到，无穷小数的乘法会比加法更为复杂。

首先，有一个产生这个小数的很自然的办法。这个数肯定在1和2之间，因为$1^2=1$，小于2，而$2^2=4$，大于2。如果你一一计算1.1^2，1.2^2，1.3^2，一直到1.9^2，你会发现$1.4^2=1.96$小于2，而$1.5^2=2.25$大

于2。所以$\sqrt{2}$必然在1.4和1.5之间．因此它的小数展开一定是以1.4开头。现在假设你已经以此办法算出$\sqrt{2}$的前八位数字是1.414 213 5。你可以继续如下的计算，得到下一位数字是6。

1.414213502＝1.9999998236822500
1.414213512＝1.9999998519665201
1.414213522＝1.9999998802507904
1.414213532＝1.9999999085350609
1.414213542＝1.9999999368193316
1.414213552＝1.9999999651036025
1.414213562＝1.9999999933878736
1.414213572＝2.0000000216721449

重复这一过程，你想算出多少位都可以。虽然这样的计算永远无法结束，但是你至少有了一种清晰明了的方法，能给定小数点后第n位的数字，无论n取何值：小数点后有n位数，平方不大于2的最大小数，$\sqrt{2}$的小数点后第n位数字就与它的最后一位相同。例如，在所有小数点后有两位的小数中，1.41是平方不大于2的最大的一个，所以$\sqrt{2}$展开到小数点后两位是1.41。

让我们记最终得出的无穷小数为x。是什么让我们能够坚信x2＝2呢？我们可能会给出如下的理由。

12＝1
1.42＝1.96
1.412＝1.9881
1.4142＝1.999396
1.41422＝1.99996164
1.414212＝1.9999899241
1.4142132＝1.999998409469
1.41421352＝1.99999982368225
1.414213562＝1.9999999933878736

上述算式列表显示了$\sqrt{2}$的小数展开位数越多，自身相乘得到的小数点后的数字9就越多。因此，如果把$\sqrt{2}$完整地展开到无穷多位，我们应该得到无穷多个9，而1.99999999…（9无限循环）等于2。

这样的论述会导致两个困难。其一，为什么1.999999…等于2？其二，也是更严重的一个问题，“完整地展开到无穷多位”是什么意思？这是我们首先想努力搞懂的问题。

为了解决第一个困难，我们必须再次搁置一切直觉。在数学中，人们的确普遍认为1.999999…等于2，但这个事实并不是经由某种形而上学的推理过程发现的。相反，它只是种传统习惯。但这也绝不是毫无理由的传统，因为如果不采纳它，我们就必须发明怪异的新对象，或者抛弃一些熟悉的算术规则。比如，如果你坚持1.999999…不等于2，那么2-1.999999…是什么呢？如果这是0，那么你也就抛弃了一条有用的规则：若x-y＝0 ，则x必定与y相等。如果这不是0，那这个数就没有通常的小数展开（否则，你从2中减掉这个数，得到的就不是1.999999…，

而是较小的别的数），所以你必须创造一个新的对象，诸如“0加小数点后无穷多个0，之后是个1”。如果做这样的事，那你的麻烦才刚刚开始。这个神秘的数自乘是什么东西呢？小数点后无穷多个0，之后再来无穷多个0，之后是个1？如果你用它乘以10呢？你得到的是不是“无穷多减1”个0，之后是个1？1/3的小数展开是什么呢？用它乘以3，答案是1还是0.999999…？如果你遵循了惯用的传统，这一类难以对付的问题就不会产生了。（虽难以对付，但也并非全然不可能：在1960年代，亚伯拉罕·罗宾逊发现了对“无穷小”数的一个条理清晰的定义，但正如其名“非标准分析”所暗指的，这套理论还没有成为数学界的主流部分。）

第二个困难是更实实在在的困难，但它也是可以绕过去的。我们不去试图想象对无穷小数进行长乘计算，而仅仅将陈述$x^2=2$解释为，x展开位数越多，平方的结果就越接近2，一如我们已经观察到的。为了更确切地说明这一点，假设你仍然坚持想要一个数，它的平方以1.9999…开始。我会向你建议1.414 21这个数，它是由x的前几位给出的。因为1.414 21很接近1.414 22，我预期它们的平方也会很接近（这很容易证明）。但根据我们选择x的办法，必然有$1.414\ 21^2$小于2而$1.414\ 22^2$大于2。因而这两个数都很接近于2。只要检查一下：$1.414\ 21^2=1.999\ 989\ 924\ 1$，于是我们就找到了满足条件的一个数。如果你想要一个新的数，其平方开始几位是

1.99999999999999999999999999999999999999999999999…

那我可以用同样的办法，只不过多取x几位而已。（事实上，如果你想要n个9，那取到小数点后n+1位就足够了。）无论你想要多少个9，我都可以办到，这一事实正是无穷小数x自乘等于2的含义。

请注意，我所做的是“驯服”无穷，只是将涉及无穷的陈述单纯解读为一种生动的简化，其所指的乃是一条不涉及无穷的累赘得多的陈述。关于无穷的简洁陈述是“x是平方等于2的无穷小数”。可以大致翻译成：“有这样一种规则，对任意n，它能够切实地给出x的前n位数字。这使我们能够算出任意长的有限小数，它们的平方接近于2，只要算得足够长，想要有多接近就能有多接近。”

我是不是在说，$x^2=2$这个外表简单的陈述的真正意思其实非常复杂呢？某种意义上我的确是这个意思——这条陈述确实隐含了复杂性；但从更重要的意义上讲，我并非是指这个意思。固然，在不提及无穷的情况下，要定义无穷小数的加法和乘法是很难的，而且我们还必须检查这个复杂的定义遵从第二章中列出的那些规则，诸如交换律和结合律。但是，一旦给出了这个定义，我们就可以再次无拘无束地进行抽象思考。关于x，重要的是它的平方等于2。关于“平方”一词，重要的是它的含义以某种乘法定义为基础，这种定义遵循着恰当的规则。x的第一万亿位是什么并不是真正紧要的，乘法的定义有多复杂也不真正紧要。

2. 我们经过那个路灯柱时时速为40英里

假设你坐在一辆正在加速的小轿车里，看着速度表的指针稳定地从时速30英里移动到时速50英里。我们很容易说，在某个时刻——恰在指针扫过40的那一瞬间——汽车以时速40英里的速度行驶。在那个时刻以前，车速小于它，在之后车速大于它。但是，说仅在某个瞬间车速为每小时40英里，这是指什么意思呢？如果汽车没有加速，那我们可以测量它一小时走过多少英里，从而得到它的速度。（更具操作性的是，观察它在30秒内走过多远，再乘以120。）但是，这种办法对加速的汽车显然无效：测量它在一段时间内走过多远，只能算出这段时间内的平均速度，并不能得到特定瞬时的速度。

如果我们能够测量在无穷小的一段时间内汽车走了多远，那问题就不存在了，因为这样速度就没有时间发生变化。如果这段时间是t小时，t是个无穷小的数，那么我们可以测量汽车在t小时走过多少英里，记测量结果为s，它当然也是一个无穷小的数，用它除以t就得到了汽车的瞬间速度。

这种荒唐的幻想引起的问题相当类似于前面遇到的情况，当时我们简要地考察了1.999999…可能不等于2的思想。t是0吗？如果是，那很明显，s必然也是0（没有时间流逝，汽车不能走过任何路程）。但是，无法通过0除以0来得到清楚明白的答案。如果t不是0，那在这t小时之内汽车就会加速，从而导致测量无效。

要想理解瞬时速度，我们就要利用这样的事实：如果t非常小——比如百分之一秒，则汽车没有时间去加速太多。设想对于某个时刻，我们不去试图计算确切的速度，改而用一个较准的估计来代替。这样，如果我们的测量装置是精确的，我们就能知道汽车在百分之一秒内走了多远，然后把这段距离乘上一小时内百分之一秒的数量，即乘以360 000。这个答案并不会特别准确，但既然汽车在百分之一秒内加速不多，这个答案就给了我们一个相当接近的近似值。

这个情况让我们回想起1.414 213 5^{2}相当近似于2这一事实，于是类似地，我们可以不用再操心无穷，或者说是本例中的无穷小。假设我们不是在百分之一秒内测量，而改为在百万分之一秒内测量，那汽车的加速就更少了，因而得到的答案会更加精确。这种观察让我们有了一种办法，可以把陈述“汽车此刻的行驶速度是40英里每小时”转换成以下这条更复杂的、不涉及无穷的陈述：“如果你给定了允许的误差范围，那么只要t足够小（远小于1），我就可以测出汽车在t小时内走过的英里数，再除以t，得到的结果将很接近于40英里每小时，误差在你规定的范围之内。”例如，如果t足够小，我可以保证使我的估计值位于39.99和40.01之间。如果你要求我的答案精确到误差小于0.0001，那么我可以使t变得更小。只要它足够小，我总可以满足你要求的精度。

我们再一次把一条涉及无穷的陈述，视为对一条更复杂的、关于近似的命题的方便表达。另一个更具提示性的词语是“极限”。一个无穷小数是一列有限小数的极限，瞬时速度是通过测量越来越短的时间内走过的距离所得近似值的极限。数学家经常谈论“在极限时”或者“在无穷时”的情况如何，但他们都很明白，他们并没有把这种说法完全当真。如果强迫他们说出确切意思，他们就会转而谈论近似。

3. 半径为r的圆面积为πr^2

无穷可以借由有穷的情况来理解——人类对这一点的认识就是19世纪数学的伟大胜利之一，不过这种思想的根源能够追溯到更早。我的下一个例子是讨论圆面积的计算，就将使用阿基米德在公元前3世纪所提出的论证。不过，在开始计算以前，我们应当明确要算的是什么，这个问题并没有乍看上去那么简单。什么是面积？当然，它是图形里某种东西的量（二维的东西），但怎样才能精确表达呢？

不论它是什么，对特定形状来说，它看起来当然是很容易计算的。例如，如果矩形的边长为a和b，那它的面积就是ab。任何直角三角形都可以被看作把矩形沿一条对角线切下一半所得，因此它的面积是对应矩形面积的一半。任意三角形都可以分割成两个直角三角形，任意多边形都可以分割成一系列三角形。因此，要算出多边形的面积并不会太难。不必去烦恼我们算的到底是什么，我们只要单纯地将多边形的面积定义为此计算结果即可（只要我们能够确信，将多边形切割成三角形的不同方法不会导致不同的答案）。

当我们考虑边缘为曲线的图形时，问题就来了。将圆切割成若干个三角形是不可能的。那么当我们说圆的面积是πr^2时，我们谈的是什么呢？

这又是抽象方法大有用处的一个例子。让我们不要关注面积是什么，而是关注面积能够做什么。我们要对这个说法作一些澄清，因为面积看上去做不了什么事情——它只不过是就在那里而已。我所说的意思是，我们应当关注，关于面积的任何合理概念所应具有的一些属性。下面列出了五条。

Ar1 移动图形，图形面积不变。（更正式的说法：两个全等的图形面积相等。）

Ar2 如果一个图形完全包含于另一个之内，那么第一个的面积不大于第二个。

Ar3 矩形的面积通过它的长宽相乘得到。

Ar4 将图形切成若干部分，则各部分面积之和等于原图形的面积。

Ar5 图形向各方向扩张为原来的2倍，则图形面积变为原来的4倍。

如果回过头来看，你会发现我们利用了属性**Ar1**、**Ar3**和**Ar4**来计算直角三角形的面积。属性**Ar2**看起来非常显然，几乎无须一提，不过作为公理正应如此，而且我们后面还会看到它会多有用。属性**Ar5**尽管很重要，其实并不需要作为一条公理，因为它能从另外几条中推导出来。

我们怎样利用这些属性来谈论圆面积的含义呢？截至目前，这一章所要向读者传达的信息就是，不要考虑一步到位的定义，考虑面积的近似可能会有收获。我们可以采取下面这种办法轻易做到。想象一个图形画在一张布满细密方格的纸上。由属性**Ar3**，我们知道这些方格的面积（因为正方形只是特殊的矩形），所以我们可以数一数有多少正方形完全处于图形以内，以此来估计图形的面积。举例来说，图形中包含了144个小方格，那么图形的面积至少是144乘以单个方格面积。注意，我们所真正计算的面积仅是144个方格所构成的图形的面积，此面积由属性**Ar3**和**Ar4**可以很容易地确定。

对于图18所示的图形，这并不能给出正确的答案，因为还有若干个方格部分位于图形之内、部分位于图形之外，我们还没有计入它们的面积。不过，有一种改进估值的明显方法，就是把每一个小方格分成四个更小的方格，改用它们来计算。和之前一样，又有一些方格部分位于图形外、部分位于图形内，但这回我们已经多包含进了一些完全在图形之内的小方格。一般来说，方格划分得越密，我们在计算中能够计入的原图面积也就越多。我们发现（这一点并没有看上去那么显然），随着把网格画得越来越密，方格越来越小，计算的结果也就越来越接近于某个数，就像$\sqrt{2}$近似值的平方越来越接近于2一样，于是我们就将这个数定义为这个图形的面积。

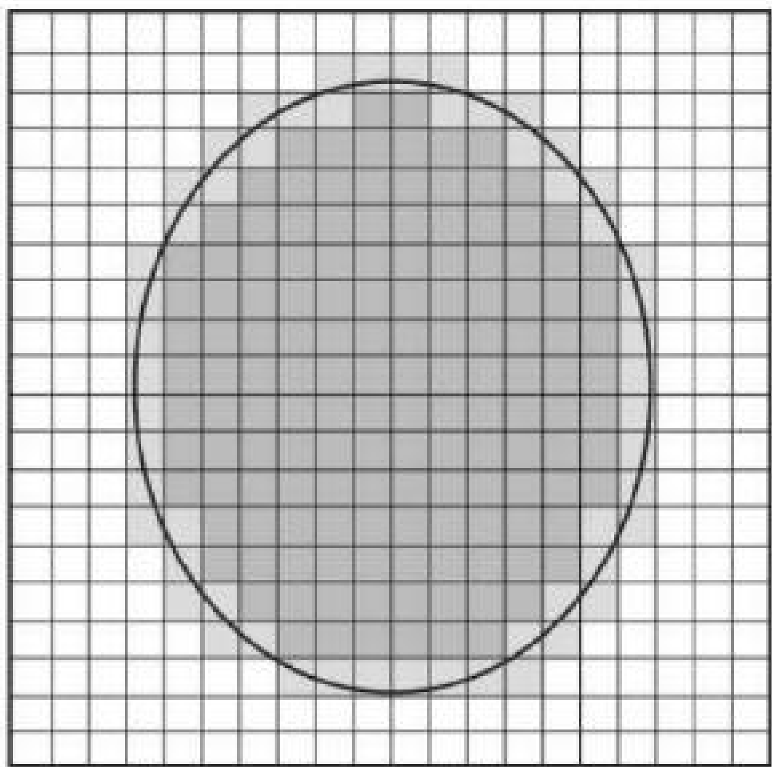


图18 曲线图形面积的逼近

所以，从数学的角度来讲，某个图形的面积为一平方码的含义是这样的：在特定的容许误差范围之内，无论多小，总可以找出充分密集的方格，由位于图形内的小方格来近似计算，得到的计算结果与一平方码之差少于给定误差。（在脑子里某处可能有这样的想法——但已经被压制住了："在极限下"可以使用无穷多无限小的方格来得到确切的答案。）

还有另外一种说明方法，也许能够说得更清楚。如果一个曲线图形的确切面积是12平方厘米，要求我用方格法来表明这一点，这个任务是不可能完成的——我需要用无穷多的方格才能做到。但是，如果你给出任何其他不同于12的数字，比方说11.9，那我都可以用一套方格来确定

地证明它的面积不是这个数：我只需使网格足够密，让没有计入的部分小于0.1平方厘米即可。换句话说，在不涉及无穷的情况下，我不去证明它的面积是12，而是满足于证明它的面积不是别的任何数。图形的面积是我所不能证伪的那个数。

这些思想给了面积一个令人满意的定义，但仍给我们遗留下一个问题。我们要怎样来说明，如果采用上述过程来估计半径为r的圆的面积，估计值就会越来越接近πr2呢？对于大多数图形来说，答案是必须使用到积分，我这本书中不会去讨论它。但对于圆来说，正如我之前所提到的，我们可以使用阿基米德的绝妙论证。

图19表示的是，将圆切成一片一片，再组合成接近于矩形的形状。因为每一小片都很窄，所以矩形的高度大约为圆的半径r。同样因为每一片很窄，这个准矩形的上边和下边都近似于直线。上下两边各用了圆的周长的一半，由π的定义知圆的周长为2πr，则两边长度均近似为πr。因此，准矩形的面积是r×πr＝πr2——至少是近似如此。

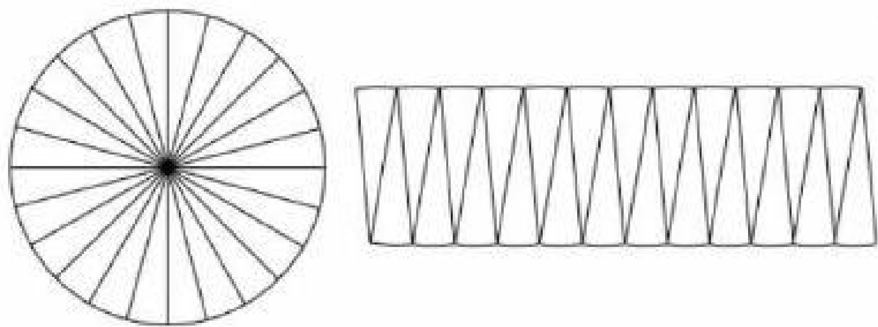


图**19** 阿基米德说明圆面积为π**r2**的方法

当然了，事实上准矩形的确切面积就是πr2，因为我们所做的只不过是把圆切开，再移动各部分而已，但是我们目前还不知道这一点。到现在为止，这个论证可能已经使你确信了，但其实还没结束，因为我们还必须表明，随着切开的片数越来越多，上述近似值会越来越接近πr2。一种非常简明的方法是利用两个正多边形，一个恰好包含在圆内，一个恰好包含圆。图20用六边形说明了这一点。内接多边形的周长小于圆的周长，而外切多边形的周长大于圆的周长。这两个多边形都可以切成多个三角形，再拼合成平行四边形。一点简易的计算就能够表明，较小的平行四边形面积小于r乘以内接多边形周长的一半，因而小于πr2。类似地，较大的多边形面积大于πr2。如果切出的三角形数量足够大，那么这两个多边形面积之差就会越来越小，想要多小就有多小。因为圆总是包含着小多边形，并被大多边形所包含，所以它的面积必然恰为πr2。

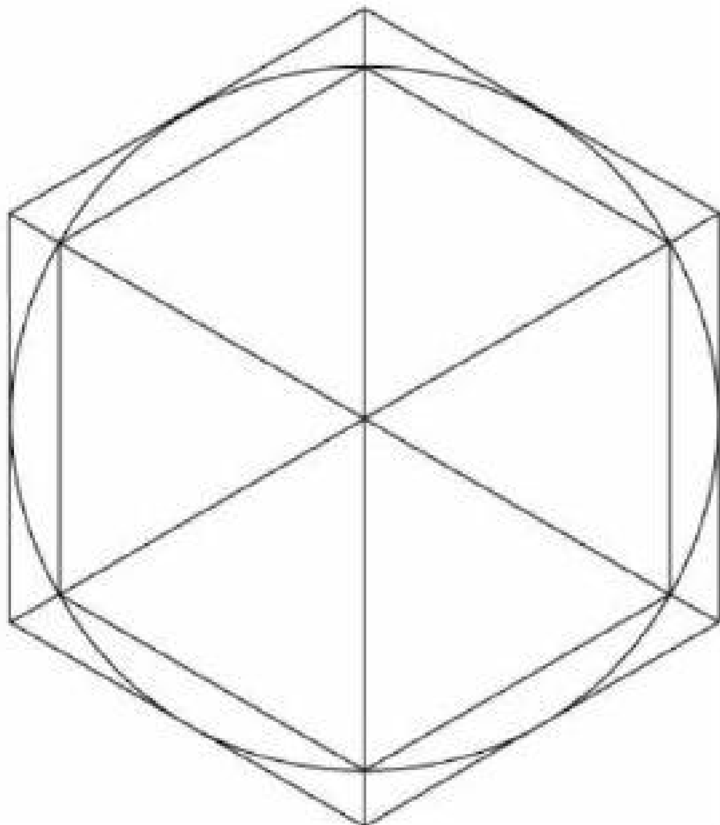


图20 由多边形逼近圆

第五章
维度

高等数学有个显眼的特征，它的大部分内容涉及高于三维的几何。这个事实令数学家以外的人感到很困惑：直线和曲线是一维的，曲面是二维的，形体是三维的，但怎么会有东西是四维的呢？一旦物体具有了高度、宽度和厚度，它就完全填充着空间的一部分，而且看起来已经没有其他维度施展的余地了。有人会提出第四维是时间——在特定情境下这是个不错的回答，比如在狭义相对论中。但它并不能帮助我们理解诸如二十六维乃至无穷维的几何，而这些几何在数学上又都很重要。

高维几何又是一例最好从抽象角度来理解的概念。让我们不去担心二十六维空间的存在等等，而去考虑它的性质。你可能会疑惑：这东西连是否存在都不确定，怎么可能考虑它的性质呢？不过这样的担心很容易解决。如果丢弃“这东西”一词，那么这个问题就变成了：连拥有这种性质的对象是否存在都不知道，怎么可能考虑这种性质呢？但这个问题一点都不困难。比如说，我们完全可以思考一位美国女总统可能会具有的特质，即便并不能保证就会有一位女总统。

我们会期待二十六维空间有什么样的性质呢？最明显的一种，即使它成其为二十六维的性质，是它通过二十六个数来确定一点，就像两个数确定二维空间中一点、三个数确定三维空间中一点一样。另一个性质是，如果你取一个二十六维形体，使它在各个方向上都扩张两倍，那它的“体积”——假设我们能够使这个词有意义——应当乘以2^{26}。此外还有其他的一些性质。

若最终发现二十六维空间的概念在逻辑上不能自洽，那这样的思考就没什么意思了。为了消除疑虑，我们终归还是要表明它是存在的——如果它不自洽，明显就不存在——不过是从数学而非物理的角度。这么说的意思是，我们需要定义一个恰当的模型。它可能并不非得是依附于某种东西的模型，但如果它具有我们所想要的全部性质，那它就表明这些性质是逻辑自洽的。不过，一如往常，结果表明我们将定义的这个模型非常有用。

怎样定义高维空间？

定义这个模型出奇地容易，只要怀有一个想法——坐标系。我前面说过，二维中的一点可以由两个数确定，三维中则需要三个数。常规做法是采取笛卡尔坐标系，之所以这样称呼是因为它是由笛卡尔发明的（他声称自己是在梦中产生这个想法的）。在二维上，先画出垂直相交的两个方向。例如，一个方向可能向右，另一个方向径直向上，如图21所示。给定平面上任意一点，你都可以通过水平移动一段距离（如果是向左移动，则认为是向右移动了负的距离），再转90度并垂直移动另一距离到达这一点。这两段距离给了你两个数，这两个数就是你所到达的这一点的坐标。图21中显示出了坐标为（3，2）的点（右移三上移二），坐标为（-2，1）的点（左移二上移一）以及坐标为（1，-2）的点（右移一下移二）。完全同样的程序在三维，也就是在空间中也有效，只不过你必须使用三个方向，比如向前、向右和向上。

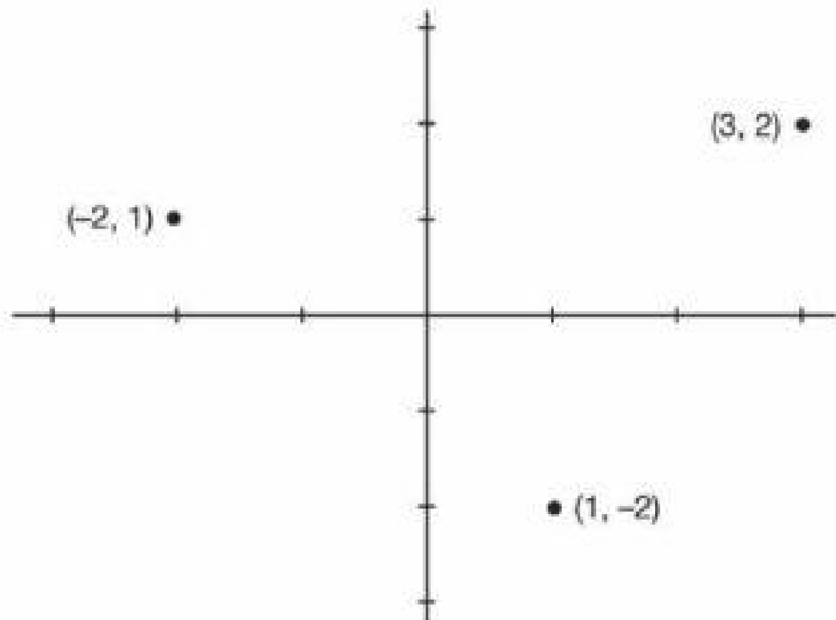


图**21** 笛卡尔平面上的三点

现在让我们稍微改变一下视角。我们不再将这两个数（或三个）称作点在空间中的坐标；让我们说，这些数就是点。也就是说，我们不再说“坐标为（5，3）的点”，让我们说“（5，3）这个点”。有人可能把这仅当作语言上的便利，但它实际上有更深的意义。它在用空间的数学模型取代实在的、物理的空间。我们的二维空间数学模型是由成对的实数（a，b）所组成的。尽管这些成对的数本身并不是空间中的点，我们也称它们为点，因为我们希望提醒自己，这正是它们所表示的东西。类似地，我们可以取所有的实数三元组（a，b，c）得到三维空间的模型，并把这些三元组称为点。现在我们就有了定义诸如八维空间中点的很明显的方法。它们只不过是实数八元组。例如，这就是两个点：（1，3，-1，4，0，0，6，7）和（5，π，-3/2，$\sqrt{2}$，17，89.93，-12

$\sqrt{2}+1$）。

我现在已经定义了一种初步的数学模型，但它还不值得被称作八维空间的模型，因为“空间”一词包含着很多几何含义，我还没有用模型来描述它们：空间不仅仅只是大量单个点的堆砌而已。例如，我们会谈论一对点之间的距离，会谈论直线、圆及其他几何形状。这些思想在高维空间中的对应物是什么呢？

回答这类问题，有一个通用的方法。找出一个二维或三维中熟悉的概念，首先完全用坐标的语言来描述它，然后就能预期，它向高维空间中的拓展变得很显然。让我们看看这个方法是怎么处理距离这一概念的。

给定平面上两点，如（1，4）和（5，7），我们可以按如下方法计算它们的距离。首先按图22所示画一个直角三角形，其另一个顶点位于（5，4）。我们注意到，连结（1，4）和（5，7）的线段是这个三角形的斜边，这意味着可以用毕达哥拉斯定理来计算出它的长度。另两条边的长度为5-1＝4和7-4＝3，所以斜边的长度是 $\sqrt{4^2+3^2}=\sqrt{16+9}=5$。因此两点之间的距离为5。将这种方法应用到一般的一对点（a，b）和（c，d）上，我们得到一个直角三角形，其斜边端点正是这两个点，另两条边的长度为|c-a|（这表示c和a之间的差距）和|d-b|。毕达哥拉斯定理告诉我们，两点间的距离由下式给出：

$$\sqrt{(c-a)^2+(d-b)^2}$$

类似的方法在三维中也有效，只不过稍微复杂一点，可以得出（a，b，c）和（d，e，f）两点间的距离是：

$$\sqrt{(d-a)^2+(e-b)^2+(f-c)^2}$$

换言之，要计算两点间的距离，你需要将对应坐标差的平方相加，然后求平方根。（简要地说，理由如下：以（a，b，c）、（a，b，f）和（d，e，f）为顶点的三角形T，在点（a，b，f）处是个直角。（a，b，c）到（a，b，f）的距离是|f-c|，而由二维公式，（a，b，f）到（d，e，f）的距离是 $\sqrt{(d-a)^2+(e-b)^2}$。在T内应用毕达哥拉斯定理，就能得到结果。）

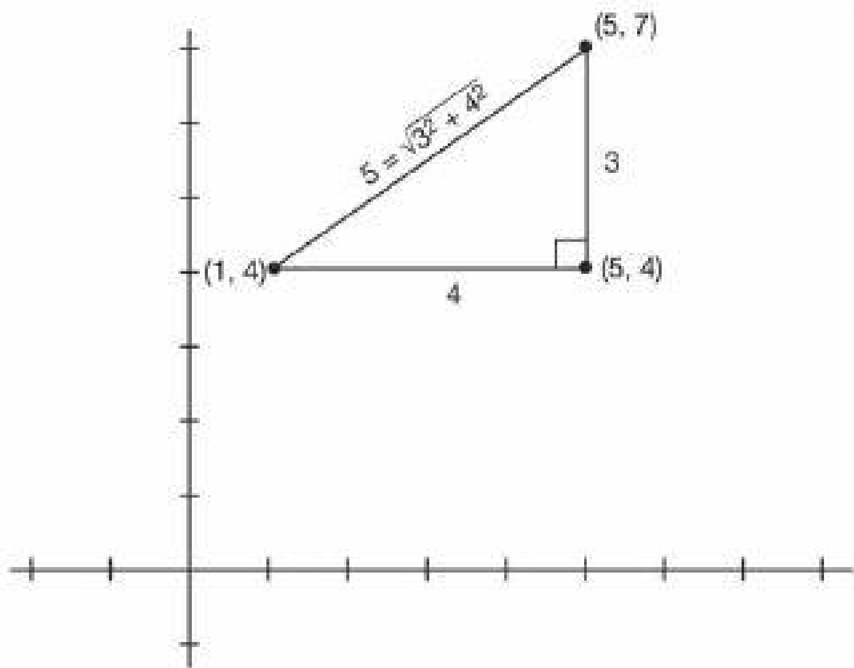


图22 用毕达哥拉斯定理计算距离

这条陈述有个有意思的特征：它没有提到假设点在三维空间中这个事实。因此我们凑巧发现了计算任意维空间中距离的方法。例如，（1，0，-1，4，2）和（3，1，1，1，-1）（五维空间中）这两点间的距离是：

$$\sqrt{(3-1)^2+(1-0)^2+(1-(-1))^2+(1-4)^2+(-1-2)^2}=$$

$$\sqrt{4+1+1+9+9}=\sqrt{27}$$

这种处理方式有一点误导作用，它暗示着任意一对五维空间中的点之间总是有一个距离（要记住，五维空间中的点只不过意味着五个实数而已），而我们发现了怎样把这个距离计算出来。但实际上，我们所做的是定义距离的概念。没有什么物理实在强迫我们必须要按上述方法计算五维空间的距离。但另一方面，这种方法很明显，就是我们在二维和三维空间所用方法的自然拓展，若要采用别的定义则会显得很奇怪。

一旦定义了距离，我们就可以开始扩展其他概念。例如，球面显然就是圆在三维空间中的等价物。那四维空间中的“球面”会是什么呢？和距离一样，如果我们把二维和三维版本的概念用一种不提及维数的方法描述出来，就可以回答这个问题了。这其实根本不难：圆和球面都可以通过到定点（圆心或球心）固定距离（半径）的所有点的集合来描述。在此方面，我们完全可以把相同的定义用于四维中的球面，甚至八十七维中的球面。例如，四维空间中以（1，1，0，0）为定点以3为半径的球面，正是所有到（1，1，0，0）的距离为3的（四维的）点所组成的集合。四维空间中的点就是四个实数（a，b，c，d）。它到（1，1，

0，0）的距离是（根据我们之前的定义）：

$$\sqrt{(a-1)^2+(b-1)^2+c^2+d^2}$$

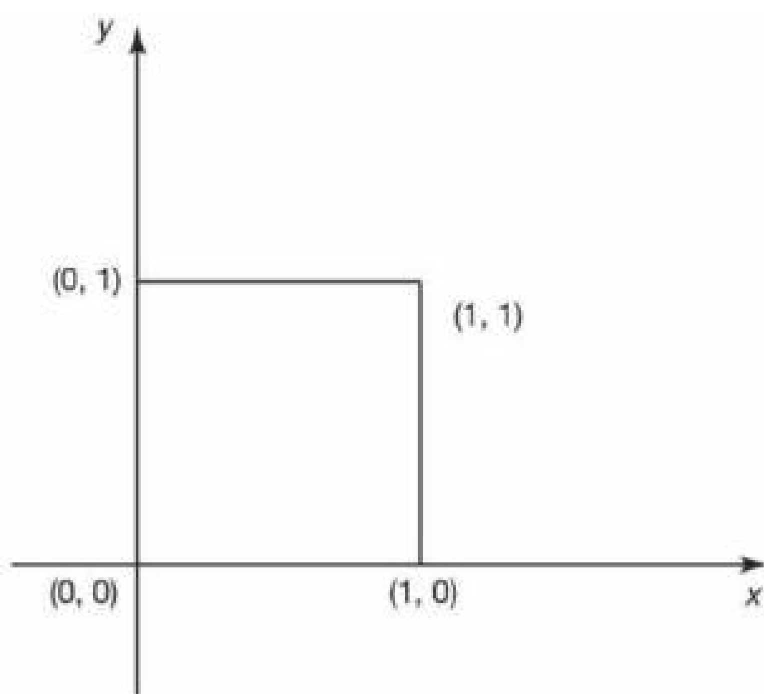
因此，描述这个四维球面的另一种方式是：它是满足如下条件的所有四元组（a，b，c，d）的集合。

$$\sqrt{(a-1)^2+(b-1)^2+c^2+d^2}=3$$

例如，（1，-1，2，1）就是这样的一个四元组，因此它是这个四维球面上的一点。

另一个可以扩展的概念是二维中的正方形和三维中的立方体。如图23所示，满足a、b均在0和1之间的所有点（a，b）的集合就形成了边长为1的正方形，它的四个顶点是：（0，0），（0，1），（1，0）和（1，1）。在三维中，我们可以取满足a、b、c均在0和1之间的所有点（a，b，c）的集合，以此来定义一个立方体。它的八个顶点是：（0，0，0），（0，0，1），（0，1，1），（0，1，0），（1，0，0），（1，0，1），（1，1，0）和（0，0，0）（1，1，1）。显然我们可以在高维空间中采用类似的定义。比如，我们可以通过取各个坐标均在0和1之间的所有点（a，b，c，d，e，f）的集合，以此得到六维立方体，或者说一个明显当得起这个名字的数学构造。它的顶点就是所有那些坐标非0即1的点。不难看出，维数每增加1，顶点的个数就翻倍，所以这个例子中顶点共有64个。

除了单纯定义形状之外，我们还可以做许多其他事情。为了简要说明这一点，我来计算一下五维立方体的边的条数。我们并不能一下子就看出边是什么，但我们可以从二维和三维的情形中得到一些线索：边就是将相邻两个顶点连结起来的线段，两个顶点如果只有一个坐标不同，那就认为它们是相邻的。取五维立方体的一个常见顶点，如（0，0，1，0，1），根据刚刚给出的定义，它的相邻顶点是（1，0，1，0，1）、（0，1，1，0，1）、（0，0，0，0，1）、（0，0，1，1，1）和（0，0，1，0，0）。一般地，每个顶点有五个相邻顶点，因而也就引出五条边。（怎样把连结相邻顶点的线段的概念从二维和三维扩展到五维，我把它留给读者自己思考。对于本次计算，它并不重要。）因为总共有2^5＝32个顶点，所以好像有32×5＝160条边。但是我们把每条边计算了两次——从它的两个端点处各算了一次，所以正确的答案是160的一半，即80条边。


y
(0, 1)
(1, 1)
(0, 0)
(1, 0)
x

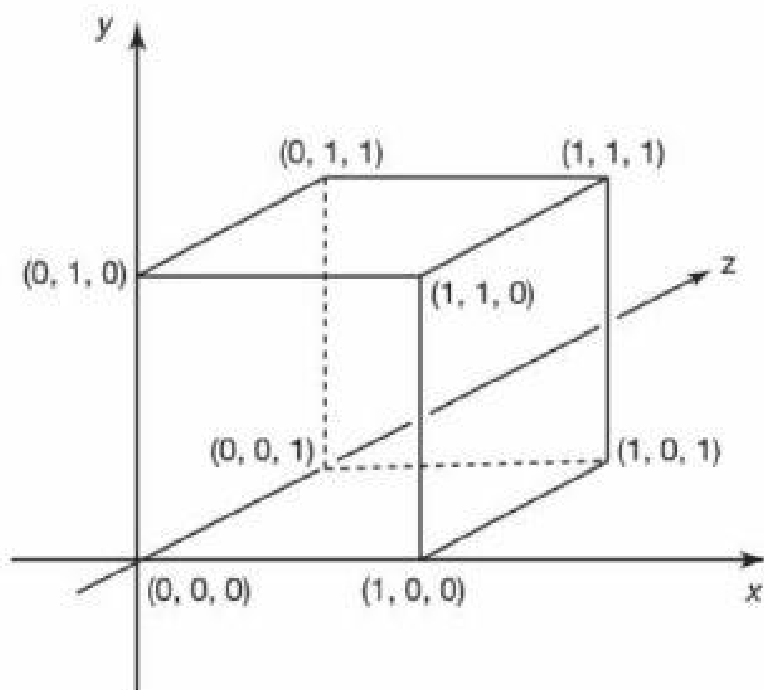

y
(0, 1, 1)
(1, 1, 1)
(0, 1, 0)
z
(1, 1, 0)
(0, 0, 1)
(1, 0, 1)
(0, 0, 0)
(1, 0, 0)
x

图23 单位正方形和单位立方体

总结上述做法的一种方式是，称之为把几何转化为代数，通过坐标系来把几何概念翻译为等价的但只涉及数之间关系的概念。尽管我们不能直接对几何进行拓展，但我们可以对代数进行拓展，而且将这种拓展称为高维几何似乎是合情合理的。五维几何显然不像三维几何那样与我们的切身经验直接相关联，但这并不妨碍我们去思考它，使它作为模型发挥作用。

四维空间能否图像化？

三维物体能够图像化而四维物体不能——这个命题看似显然，实际上却经不起严格的审视。尽管将物体图像化和直接观察它的感觉十分相像，这两种体验却有很重要的区别。例如，如果别人要求我将一个房间画出来，这个房间我比较熟悉但不是十分熟悉，这对我来说没什么困难。如果要问我一些关于它的简单问题，诸如房间里有多少椅子或者地板是什么颜色，我通常答不上来。这说明大脑中的图像不论是什么样的，还是不同于照片式的呈现。

在数学语境中，能与不能将某物图像化的重要区别在于，前一种情形下，我们能直接回答问题而无须停下来进行计算。这个“直接”当然只表示一种程度，但实际上毫不夸张。比如，如果要问我三维立方体有多少条边，我可以通过“仅仅观察”，看到顶面有4条边，底面有4条边，还有4条边从顶面伸到底面，所以一共是12条。

在高维空间中，“仅仅观察”变得困难了，我们常常被迫要像我在讨论五维空间中的推理问题时那样进行更多论证。但观察有时也是可能的。比如，我可以思考一个四维立方体，它由两个彼此相对的三维立方体组成，对应顶点以边（在第四维中）相连，就像三维立方体由两个彼此相对的正方形组成，对应顶点也连结起来一样。尽管我对四维空间没有一个完全清晰的图像，但我仍然可以“看见”两个三维立方体各有12条边，8条边连结着它们对应的顶点。这样，一共就有12＋12＋8＝32条边。于是我可以“仅仅观察”到，五维立方体又是由这样的两个四维立方体组成的，依旧是对应顶点相连，总共有32＋32＋16＝80条边（每个四维立方体有32条边，其间有16条边连结它们），恰与我之前得到的答案相同。于是，我具有了某种初步地将四维和五维图像化的能力。（如果你对“图像化”这个词感到困扰，可以换一个词，比如“概念化”。）当然，这远比三维的图像化要困难——比如，我无法直接回答，四维立方体旋转是什么样子，而三维的我就可以说出来——但是，这也明显要比五十三维的图像化要容易，要是它们都不可能的话也就谈不上谁比谁容易的问题了。有一些数学家专攻四维几何，他们四维空间图像化的能力得到了极大拓展。

对数学来说，这个心理学要素的影响已远远超出几何学的范围。投身于数学研究所能得到的乐趣之一就是，随着专业领域的经验越来越丰富，你能够发现自己“仅仅观察”就能得到越来越多问题的答案，不一定非得是几何问题，而这些问题你以前可能要艰难思考上一两个小时。举个很基本的例子，来看471×638＝638×471这条陈述。为了验证，我们可以通过两个很长的竖式乘法来计算，发现它们得到相同的结果。但是，如果考虑一个471乘638的矩形点阵，你就可以看出，第一个式子是各行点之和，第二个式子是各列点之和，所以它们必然得到相同的结果。注意，在这个问题上，我们头脑中的图像与相片化的图像是很不一样的：你真的看到了一个471乘638的矩形，而不是463乘641的矩形吗？你难道能数出短边所有的点来验证吗？

高维几何中的点是什么？

表明能够赋予高维几何某种意义是一回事，但要表明这个问题为什么值得认真对待就是另一回事了。在这一章的前面部分，我曾经说它作为模型是很有用处的。但是，既然我们所居住的实际空间是三维空间，高维几何究竟有什么用处呢？

这个问题的答案相当简单。第一章中我谈到，一个模型可以具有许多不同的功用。即使二维和三维几何也用于许多不同目的，而不仅仅是物理空间的直观模型。例如，我们表示物体的运动时，常常画一张图来记录它所走过的距离随时间的变化。这个图是平面上的一个曲线图，曲线的几何性质与物体运动的信息相对应。为什么二维几何适用于这个运动过程的模型化呢？因为在这里有两个我们关心的数——流逝的时间和走过的距离——如我所说过的，我们可以将二维空间看作所有成对的数的集合。这就提示了我们，为什么高维几何会有用处。宇宙中可能并没有潜藏着高维空间，但需要同时考虑好几个数的情形却有不少。我下面将简要描述两种情形，之后你很明显地可以发现还会有更多的类似情形。

设想我要描述一把椅子的位置。如果是向上直立着的，它的位置就完全是由两条腿与地面接触的点来确定的。这两个点可以分别通过两个坐标来描述。于是四个数就可以用于描述椅子的位置。但这四个数是有联系的，因为椅子腿底端的相互距离是固定的。如果这个距离是d，地面上两条腿位于点（p，q）和（r，s），那么由毕达哥拉斯定理有$(p-r)^2+(q-s)^2=d^2$。

这就对p，q，r，s施加了约束，我们可以用几何语言来描述这种约束：四维空间中的点（p，q，r，s）被限制在某个特定的三维“曲面”上。更复杂的物理系统也可以用类似的方式来分析，维度也变得更高。

高维几何在经济学中也很重要。例如，你如果正在犹豫买某个公司的股票是否明智，那么能帮助你进行决策的大多数信息都是以数字的形式出现的——劳动力规模、各种资产的价值、原材料的成本、利率，等等。作为一个序列，这些数可被看作某种高维空间中的一个点。通过分析许多类似的公司，你可能会确定出空间中的某个区域，认为购买此区域中的股票是不错的主意。

分数维

讨论到现在，如果说有什么事情看起来很明显的话，那就是任何形状的维数总是一个整数。你要是说需要两个半坐标来确定一个点——即便是个数学的点，这会是什么意思呢？

这个说法看起来很有道理，但我们遇到的这个困难和在第二章中定义$2^{3/2}$變这个数之前遇到的困难非常类似，当时是用抽象方法绕过了它。我们能否对维度做相似的事情呢？要想这么做，我们就必须找到与维度相关的某些性质，它和整数并没有直接的关系。这样就把与坐标数字相关的一切排除在外了，因为坐标看起来和维度这个概念联系实在太紧密了，让人很难思考其他东西。但是，的确还有另一种性质，在本章开头简短地提到过，它恰好给出了我们所需要的东西。

几何有一个重要特征会随维数变化，这就是当把形体沿各方向以因子t扩张时，有个规则决定形体尺寸如何变化。尺寸一词，我指的是长度、面积或者体积。在一维上尺寸变为t倍，或t^1，在二维上尺寸变为t^2倍，在三维上变为t^3倍。因此，t的指数就告诉了我们形体的维数。

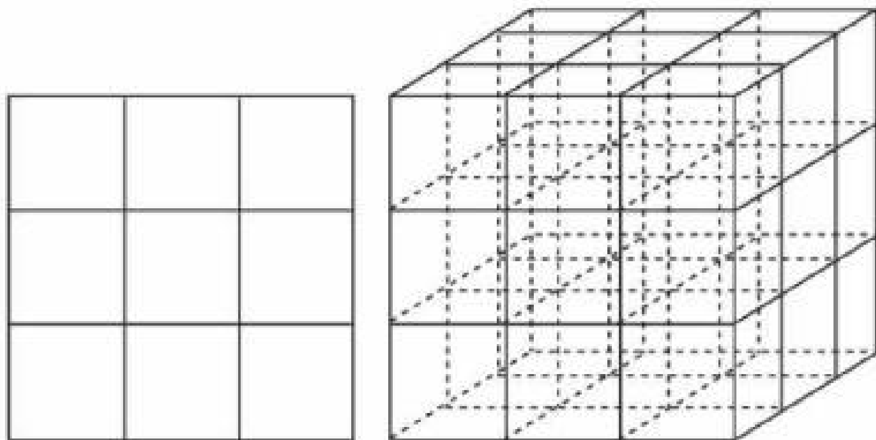


图**24** 将正方形分成**9＝32**个小正方形，将立方体分成**27＝33**个小立方体

到现在为止，我们还没有完全从描述中丢弃整数，因为数字2和3本身就隐含在“面积”和“体积”这两个词中。我们还可以不用这两个词，而采取如下的办法。为什么边长为3的正方形，面积是边长为1的正方形的9倍？原因是，我们可以把大正方形切成9个与小正方形全等的部分（如图24）。类似地，3乘3乘3的立方体，可以分成27个1乘1乘1的立方体，所以体积是小立方体的27倍。于是，我们可以说，因为当立方体以因子t扩张时（其中t是大于1的整数），则新的立方体可以分成t^3个原来的立方体，所以立方体是三维的。注意，“体积”一词没有在上一句话中出现。

现在我们可能会问：像上面一样推理得到的却不是整数，有这样的形体吗？答案是有的。一个最简单的例子就是科赫雪花。我们无法直接描述它，而是要把它定义成下列过程的极限。首先从一条直线段，比如长度为1的线段开始。把线段分成三段相等的部分，以中间一段为底，作一等边三角形，再用三角形的另外两条边来代替中间这一段。结果得到的图形是由四条直线段组成的，每一段的长度都是三分之一。再将每

条线段都切成三段相等的部分，仍将所有的中间一段都用等边三角形另两边来替代。现在的结果是16条线段，每条线段的长度是九分之一。接下来如何继续这一过程就很清楚了。开始的几步如图25所示。不难严格证明，这个过程将导致一个极限形状，正如图中所提示的那样。这个最终形状就是科赫雪花。（取三个这样的形状，连起来形成一个三角形，这样看起来会更像一个雪花。）

科赫雪花有一些很有趣的特征。和我们的主题相关的一个是，这个形状可以用它自身的微小版本建造出来。这仍然可以从图中看出来：它由四个小版组成，每一小版都是由完全版以因子三分之一收缩而得的。现在让我们来考虑，关于维数它告诉了我们什么。

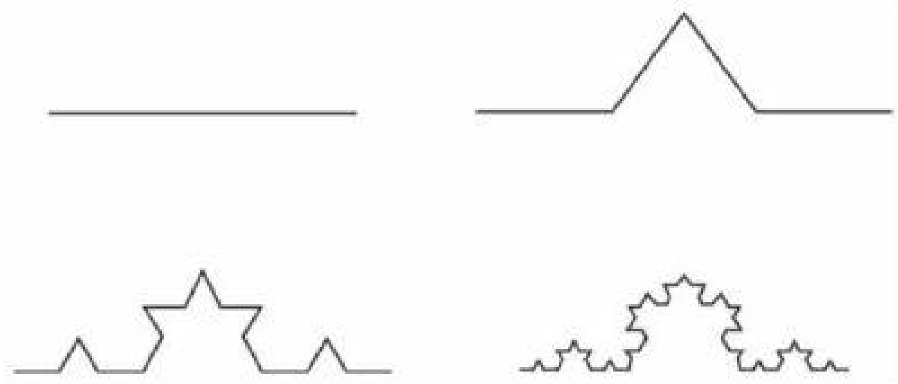


图25 建造科赫雪花

如果一个形体是d维的，那当它以因子三分之一收缩时，它的尺寸会除以3^d。（如我们所见，当d为1，2，3时，这是正确的。）这样，如果我们能够用图形的微小版本来构建出原图形，那么我们就需要3^d个小版本。因为对于科赫雪花来说需要四个，所以它的维数d应当满足$3^d=4$。由于$3^1=3$而$3^2=9$，这意味着d介于1和2之间，所以并不是一个整数。实际上，这个数是$\log_3 4$，约为1.261 859 5。

这个计算依赖于科赫雪花能够分解为较小的自身这个事实，这是个十分罕见的特征：连圆都不具备这样的特点。不过，我们还可以进一步发展上述思想，给出更为普适的维度定义。正如我们应用抽象方法的其他情况一样，这并不意味着我们发现了科赫雪花以及类似奇异形体的“真正的维度”，我们仅仅是找到了与特定性质相容的唯一可能的定义而已。实际上，还有其他定义维度的方法，会对这个问题给出不同的答案。例如，科赫雪花的“拓扑维数”是1。粗略地讲，这是因为它像直线段一样，删掉内部任何一个点后就分解成为两个不相连的部分。

对于抽象化和一般化这两个孪生的过程，这个例子提供了一点有趣的启示。我已经提到过，要将一个概念一般化，我们应当先找出与其相联系的一些性质，再将这些性质进行一般化。这样做通常只有一种自然的方式，但有时，不同的性质组合会导致不同的一般化，而多种一般化方法有时会硕果累累。

第六章
几何

亘古以来，最具影响力的数学书大概要数约公元前300年欧几里得的《几何原本》了。尽管欧几里得生活在两千多年以前，但在多种意义上，他都可以算得上第一位现代数学家——至少是我们所知道的第一位。尤其是，他是第一位系统地使用公理化方法的作者，以五条公理开始全书，并由它们推导出许许多多的几何定理。大多数人熟悉的几何——如果他们的确熟悉某种几何的话，就是欧几里得几何。但从研究的层次来讲，“几何”一词则有着更广泛的定义：当今的几何学家已经不大用到直尺和圆规了。

欧几里得几何

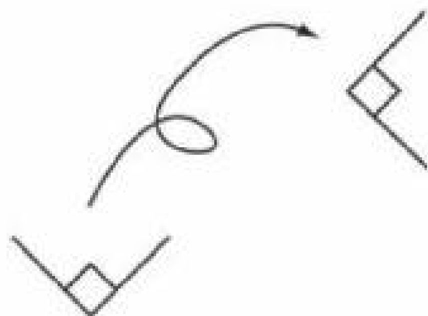
下面是欧几里得的公理。根据通常的惯例，我用“直线”一词表示两端都能无限延伸的线。“直线段”则表示有两个端点的线。

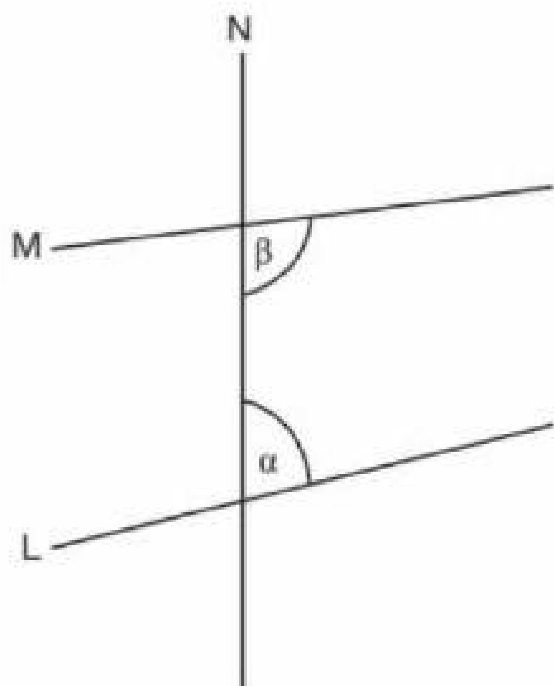
1. 任意两点有且只有一条直线段相连。
2. 任意直线段可以两端延伸形成一条直线，且只能形成一 条直线。
3. 给定任意一点p及任意长度r，存在以r为半径、p为圆心的圆。
4. 任意两个直角全等。
5. 直线N与两条直线L和M相交，若N的同旁内角之和小于两直角，则L和M相交于N的这一侧。

第四条和第五条公理如图26所示。第四条公理的意思是，如果你将一个直角移动到另一个直角上，它们恰好能够重合。对于第五条公理，由于图中所标的琢与β两角之和小于180度，所以我们知道直线L和M会在N右侧的某处相交。第五条公理与所谓的“平行公设”是等价的。平行公设断言，给定任意直线L和直线外一点x，有且只有一条直线M经过x且永远不与L相交。

欧几里得用这五条公理建造起了全部的几何学，在那之后人们也正是如此去理解几何的。例如，下面就给出了三角形内角和等于180度这个著名结果的证明框架。第一步要表明，如果直线N与两平行线L和M相交，则内错角相等。也就是说在类似图27的情况下，必定有α＝α′及β＝β′。这是第五公理的推论。首先，第五公理告诉我们α′＋β至少是180度，否则L和M就会相交（图中直线N左侧的某处）。由于α和β在一起形成一条直线，β＝180-α，有α′＋（180-α）至少是180，也就是说α′不小于α。由同样的论证，α＋β′＝α＋（180-α′）至少是180，所以α不小于α′。于是只可能是α与α′相等。又因为β＝180-α和β′＝180-α′，于是推出β＝β′。

现在考虑三角形ABC，记顶点位于A、B、C的三个角分别为α、β、γ。由第二公理，我们可以将线段AC延伸为直线L。平行公设告诉我们经过β有一条直线M与L不相交。取α′和γ′为图28所示的角。很明显有α′＋β＋γ′＝180，因为α′、β和γ′形成一条直线。因此α＋β＋γ＝180，正是我们所要求证的结论。




N
M
β
α
L

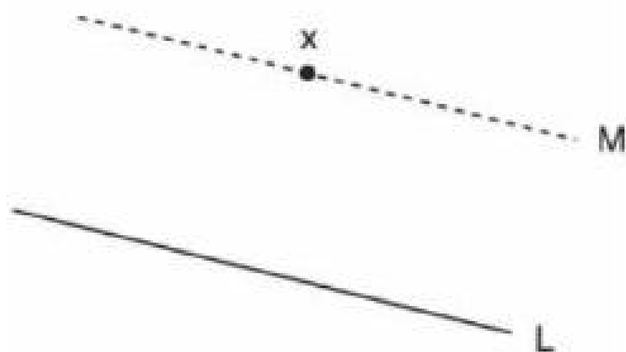

x
M
L

图26 欧几里得第四公理，及第五公理的两个版本

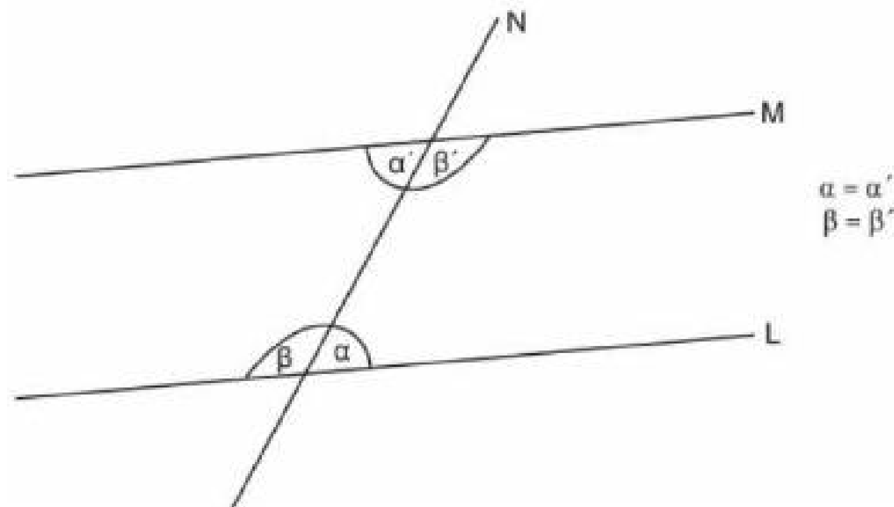


图27 欧几里得第五公理的推论

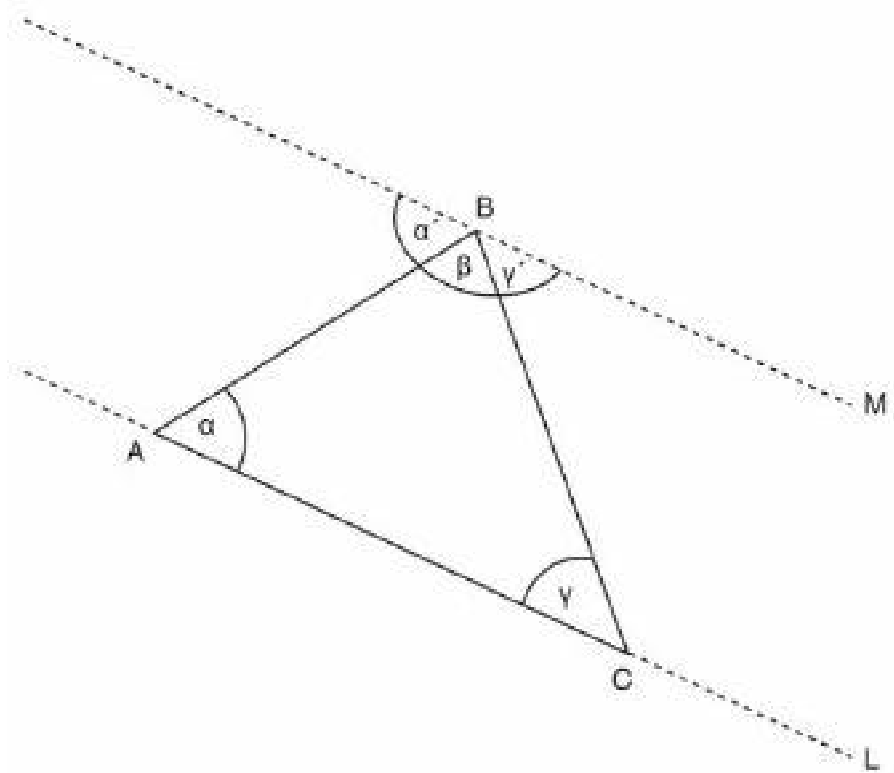


图28 三角形内角和为180度的证明

关于日常生活，这个论证告诉了我们什么呢？一个显然的结论似乎是，如果你取空间中A、B、C三点，仔细测量三角形ABC的三个内角，那它们加起来的和就是180度。一个简单的实验就可以证实这一点：在纸上画个三角形，把它尽可能整齐地剪下来，撕成三片，每片包含一个角，将各角拼在一起，可以看到这些角形成了一条直线。

如果你现在确信，不可能想象存在一个物理的三角形，其内角和不等于180度，那么你和历史上很多人站在了一起——从公元前300年的欧几里得到18世纪晚期的伊曼努尔·康德，其间所有人都得到了这个结论。康德非常相信这个结论，以至于他在《纯粹理性批判》中专门用一个显要的部分来谈论怎样绝对确定欧几里得几何为真。

但是康德错了：大约30年后，伟大的数学家卡尔·弗雷德里希·高斯就能够设想这样一个三角形。有此设想后，他实地去测量了三座山峰构成的三角形的角，分别是汉诺威王国的霍亨哈根山、因瑟尔山和布罗肯山，以检验其内角和是否为180度。（这个故事很有名，但诚实的美德迫使我在此注明，人们对于他是否真的想要检验欧几里得几何有一些怀疑。）他的试验不够具有说服力，因为要足够精确地测量这几个角很难，但这个试验的有趣之处与其说在于结果，不如说在于高斯竟然真的去费心尝试。我刚才给出的论证中究竟哪里可能错了呢？

其实，这个问题问得不对，因为上面的论证是正确的。但是，既然它以欧几里得五条公理为基础，除非这些公理在日常生活中正确，它就不能提示关于日常生活的任何事情。因此，要怀疑的是论证的前提条件，也就是要怀疑欧几里得公理的正确性。

但哪条公理看起来有一点点可疑呢？很难在它们任何一个中找出错误。如果你想在现实世界中用一条直线段连接两个点，那你只需拉紧这条线，使它通过两个点。如果你想把这条直线段延长为直线，那你可以改用激光柱来做到。类似地，画出任意半径和圆心的圆似乎也没有什么难度。经验还告诉我们，拿起纸的两个直角，确实可以使它们恰好重合在一起。最后，有什么能够阻止两条直线像无穷长的铁轨一样一直延伸到无限远？

平行公设

在历史上，引起最多怀疑的——或者至少是最让人感到不放心的公理，就是平行公设。它比其他公理都复杂，并且其基础就涉及到了无穷。当我们证明三角形内角和等于180度时，我们必须依赖于空间最远处所发生的事情，这难道不奇怪吗？

让我们更仔细地来检查一下平行公设，并且努力去理解为什么人们觉得它如此显然是正确的。或许在我们的思维深处存在着下列几项论证之一。

（1）给定直线L和直线外一点x，要通过x做一条平行线，只要选经过x并和L沿着同样的方向的线即可。

（2）记y为另一个点，它与x同在L的一侧，且到L的距离相同。用直线段连结x和y（公理1），再将这个直线段延长为直线M（公理2）。那么M就不会与L相交。

（3）记与x在L同一侧且到L距离相同的点组成的直线为M。显然它与L不相交。

以上这些论证考虑的都是平行于L的直线的存在性。下面还有一个更复杂的论证，它想要表明这样的直线至多只有一条，这是平行公设的另一个部分。

（4）将L和M用间距相同的平行线段连起来（如图29所示，就像铁轨那样），使其中一条线段经过x。现在假设N是另一条经过x的直线。在x的某一侧，N必然位于L和M之间，因此它会与相邻的一条平行线段交于点u，这个点也位于L和M之间。假设u的位置在M到L距离的1 处，那N与下一条线段交点的位置就会在2 处，依此类推。于是，经过100条线段后，N将与L相交。因为我们对N所作的假设仅仅是N不同于M而已，由此可以得出M是经过x且不与L相交的唯一的直线。

最后，对于经过一点且平行于L的直线，这里还有一个貌似能同时说明其存在性和唯一性的论证。

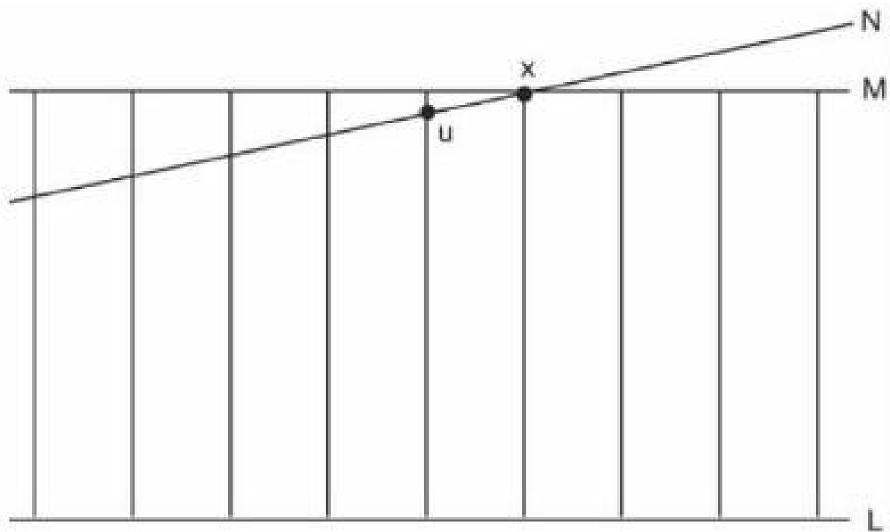


图29 平行线的唯一性

（5）平面上一点可以用笛卡尔坐标系来描述。一条（非竖直）直线L有一方程形如$y=mx+c$。通过改变c，我们可以上下移动L。显然，这样得到的直线任意两条都不可能相交，而且平面上任意一点都恰好被一条这样的直线所包含。

注意，我刚才所做的是试图证明平行公设，这也恰恰是19世纪以前许多数学家试图去做的。他们最希望将平行公设由另外四条公理推导出来，从而表明我们可以忽略掉它。然而，没有人成功过。上面我给出的这些论证及所有类似论证的毛病就在于，它们都隐含了前提假设，当把这些假设明确表达出来时，都不是欧几里得前四条公理的明显推论。尽管这些论证好像都很有道理，但它们并不比平行公设本身更有道理。

球面几何

将隐含假设明确表达出来的一个好办法，是在不同的情形下检查同样的论证。怀着这个想法，让我们来看看球面的情况。

要说平行公设在球面上不成立，这并非一望而知地显然，因为球面上根本就没有直线。我们需要使用数学中一种重要的基本思想来绕过这个难题。这种思想是抽象方法发挥作用的一个深刻例子，它就是要重新解释直线是什么意思，从而使球面上的确可以包含直线。

其实是有一种自然的定义的：一条线段就是完全位于球面内的从x到y的最短路径。我们可以把x和y想象为城市，把线段想象为飞机所走的最短路线。这样的路径必定是“大圆”，即过球心的平面切割球面所得到的圆（如图30所示）的一部分。大圆的一个例子是地球的赤道（为了讨论方便，不妨暂且把地球看作一个严格的球体）。有了定义线段的办法，大圆也就自然成了“直线”的恰当定义。

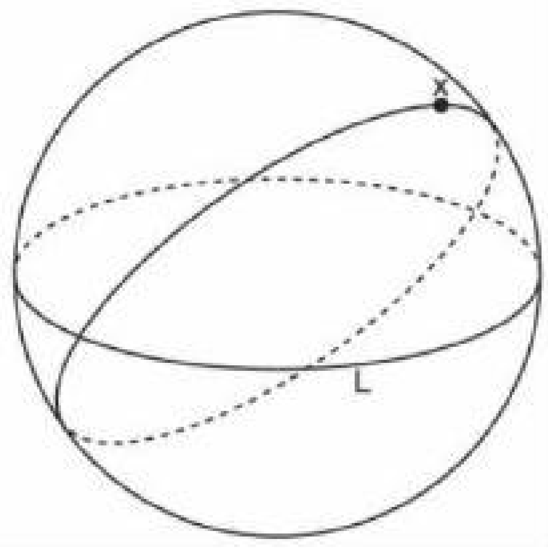
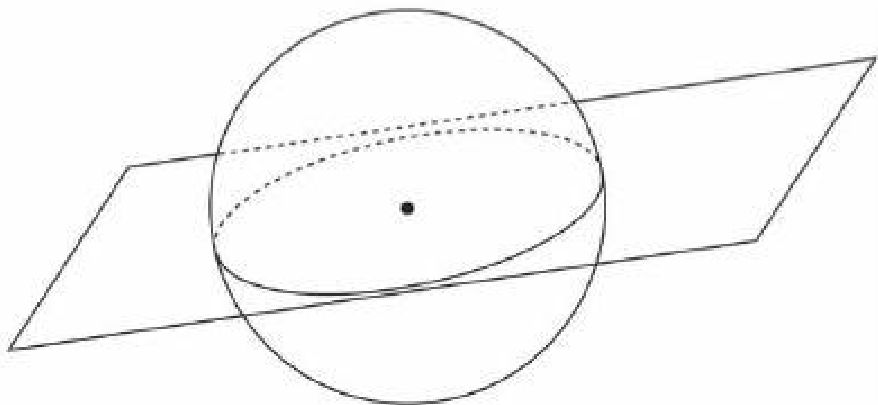


图31 球面几何中平行公设不成立

如果我们采用这样的定义，平行公设当然就不成立了。例如，设L为地球的赤道，x为北半球的一点。不难看出，任何经过x的大圆都必有一半在北半球，另一半在南半球，与赤道相交于恰好相对的两点（如图31所示）。换句话说，经过x不存在一条与L不相交的直线（我仍然指的是大圆）。


图30 大圆

这看上去好像是在耍花招：如果我用新的办法来定义“直线”，那平行公设不再成立并不特别令人吃惊。但令人吃惊不是我们的目的——实际上，我们的定义正是为了要使它不成立才如此设计的。检查一下人们在证明平行公设上的尝试会很有意思。在每个例子中，我们都能够发现在球面几何下无效的隐含假设。

比如，论证（1）假设了，“同样的方向”这个短语的意思很明显。但在球面下这一点完全不明显。为看清这一点，考虑图32所示三点，N为北极点，P在赤道上，Q也在赤道上，到P的距离为赤道的四分之一。图32中还在P的位置画了一个小箭头，沿赤道方向指向Q。那么在Q的位置上，指向同样方向的箭头应该怎么画？仍然沿着赤道方向是比较自然的。那在N处的同方向箭头呢？我们可以这样选择：从P到N画一条线段。既然P点的箭头与这条线段成直角，那N点的箭头也应该如此，那实际上也就是应当朝下指向Q。但是，我们现在就遇到了问题：我们在N点画的箭头和在Q点画的箭头指的并不是同一个方向。

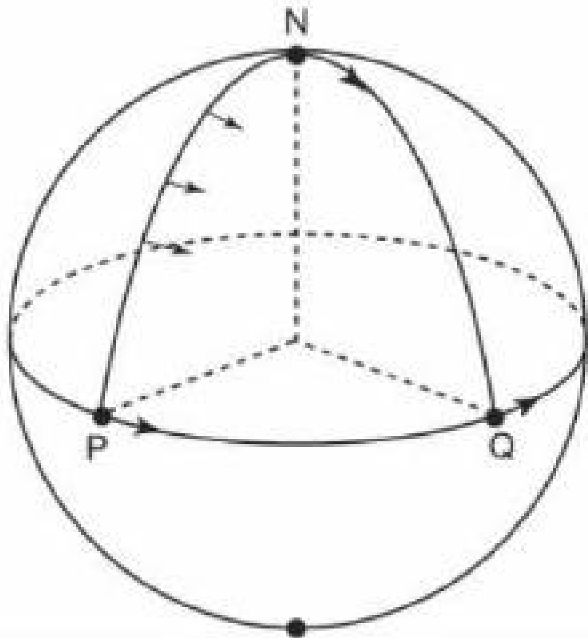


图32 球面上“同样的方向”没有意义

论证（2）的问题在于不够详细。为什么此处定义的直线M就不会与L相交？毕竟，如果L和M是球面直线的话，那它们会相交。对于论证（3）来说，它假设了M是一条直线。对球面来说这不正确：如果L是赤道，M由赤道北边1000英里的所有点组成，那么M就不是大圆。实际上，这是一条纬度恒定的线，所有飞行员和航海员都会告诉你，这不是两点间的最短路线。

论证（4）有点不一样，因为它考虑的是平行线的唯一性而不是存在性。我会在下一小节讨论它。论证（5）作出了一个巨大的假设：这个空间能够用笛卡尔坐标系来描述。但同样，这对球面来说是不正确的。

引入球面几何的意义在于，它让我们可以从论证（1）、（2）、（3）、（5）中分离出某些假设，这些假设实际上是在说：“我们所做的几何不是球面几何。”你可能会奇怪，这有什么错呢：毕竟我们做的不是球面几何。你可能还会奇怪：如果平行公设确实不是从欧几里得的其他公理中得出，我们怎么才有希望表明这一点呢？说数个世纪以来的数学家努力推导它都以失败而告终是没用的。我们怎么能确定，两百年之内会不会有某位年轻的天才能用绝妙的新思想最终得出证明？

这个问题有个漂亮的回答——至少是在原则上。欧几里得前四条公理的目的在于描述一种有限、平坦、二维空间的几何，但我们并不非要这样去解释它们——至少不是非得按照公理中那种平坦性去解释。如果我们可以将新的含义赋予“直线段”等短语，从而对公理进行重新解释（有人大概会说是“错误解释”），就像我们在球面几何中做的那样；如果我们这样做之后，发现前四条公理都是正确的但平行公设是错误的，那么我们就表明了，平行公设并非从其他公理中推出。

为了看清其中原因，可以想象一种假想的证明，从前四条公理出发，经过一系列严格的逻辑步骤，最终得出平行公设。由于这些步骤都

遵循逻辑，如果我们对其赋予新的解释，它们仍会保持有效。但在新的解释下，前四条公理都是正确的，而平行公设不正确，所以这样的论证必然是有错误的。

为什么我们不能正好用球面几何来重新解释呢？原因是，很不幸欧几里得前四条公理在球面上并不全部成立。例如，球面上不能包含半径任意大的圆，所以第三公理不成立；而且从北极到南极不止有一条最短路径，所以第一公理也不成立。所以，尽管球面几何能够帮助我们理解某些尝试过的对平行公设的证明中的缺陷，但它仍然不能保证其他可能成立的证明不存在。因此，我要转向另一种新的解释，称为双曲几何。平行公设在这里再次不成立，但这一次第一到第四公理都是成立的。

双曲几何

描述双曲几何有若干种等价的方式。我所选择的这一种称为圆盘模型，是由伟大的法国数学家亨利·庞加莱所发现的。尽管在这样一本书中我无法确切地给出定义，但我至少能够解释它的一些主要特征，并且讨论一下关于平行公设它能告诉我们什么。

理解圆盘模型比理解球面几何要复杂，因为我们不光要重新解释“直线”和“直线段”等词语，还要重新解释距离这个观念。在球面上，距离有一个很好理解的定义：x和y两点间的距离，就是在球面上从x到y最短路径的长度。尽管双曲几何中类似的定义也是正确的，但并不显然，理由涉及到什么才是最短路径——或者说任意路径的长度是什么，我们在下面能看得很清楚。

图33所显示的，是用正五边形对双曲圆盘进行镶嵌。当然我们还需要解释这句话，因为要以通常的方式来理解，图肯定是错误的：显而易见，这些“五边形”的边不是直的，长度也不相同。然而，双曲圆盘中的距离并不是用通常的方式定义的，和常规距离相比，越靠近边缘的距离越大。实际上，虽然看似不像，但边缘处的距离特别大，以至于从圆周上到中心处的距离是无穷大。所以，标星号的五边形之所以有一条边看起来比其他四条边都长，原因就是这条边最靠近中心。其他的边可能看起来短，但由双曲距离的定义，这种表面上的短正好由距离边缘的近进行了补偿。

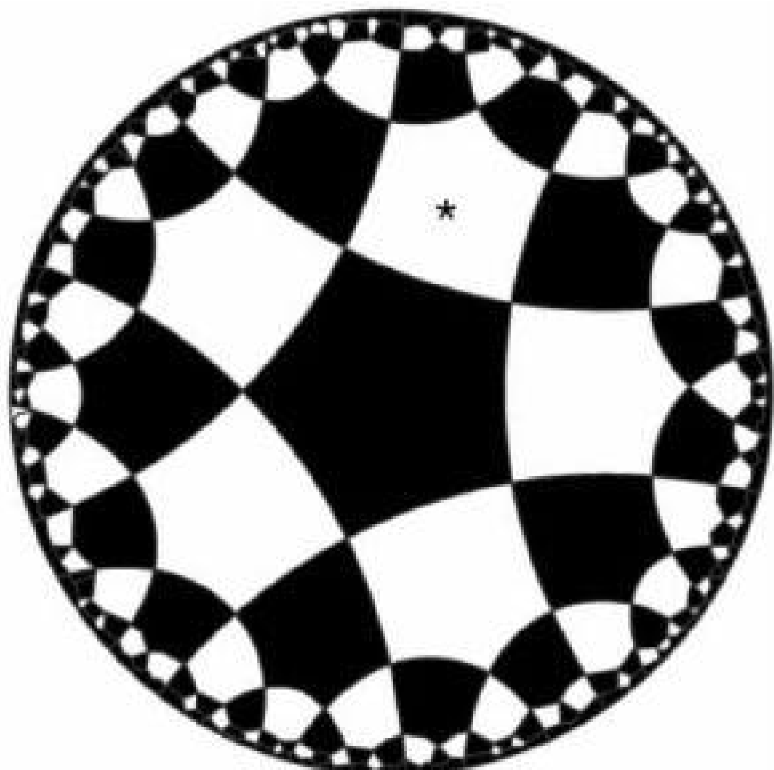


图33 用正五边形镶嵌双曲平面

如果这看起来很难解很矛盾，那不妨考虑一幅典型的世界地图。大

家都知道，因为地球是圆的而地图是平的，所以距离必然被扭曲了。表示这种扭曲的方式有很多种，最常用的一种是墨卡托投影，靠近极地的国家会显得比它们实际的要大得多。比如格陵兰就看似和整个南美洲大小相当。在这样的地图上，越靠近上端或下端，实际距离就比表面上的距离越小。

这种扭曲会产生一个众所周知的效应，地球表面上两点间最短路线在地图上就显示成了弯的。这种现象可以通过两种途径来理解。第一种是忘掉地图，想象一个地球仪，观察到如果在北半球选两点，第一点在第二点东边很远（巴黎和温哥华是个不错的例子），那么从第一点出发到第二点的最短路径会从靠近北极的地方穿过而不是伸向正西。第二种方法是直接从地图出发，根据越靠近顶部实际距离越短来推理，那么要想缩短旅程，应该同时既向西又向北走。用这种办法很难精确看出最短路径是什么，但至少“直线”（从球面距离的角度看）是弯的（从地图距离的角度看）这条原则是很清楚的。

我前面说过，越靠近双曲圆盘的边缘，与外表距离相比起来，实际距离越大。这样的结果是，两点间的最短路径倾向于朝圆盘中心偏折。这也就意味着它不会是通常意义上的直线（除非这条线恰好通过中心）。结果是，双曲直线，即双曲几何观点下的最短路径，正是与大圆边界成直角的圆弧（如图34所示）。现在再去观察图33的五边形镶嵌图，你会发现五边形的边尽管看起来不是直的，但其实都是双曲直线段，因为根据我刚才给出的定义，它们都可以延伸为双曲直线。类似地，尽管这些五边形的大小和形状看起来不都一样，但实际上的确是一样的，因为靠近边缘处的五边形要比它们看起来的大得多——与格陵兰的例子正相反。如同墨卡托投影一样，圆盘模型也是实际双曲几何的扭曲的“地图”。

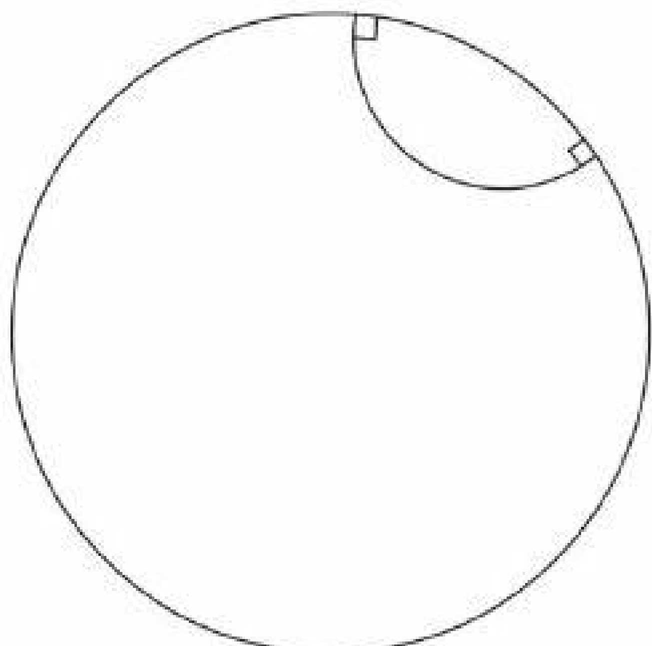


图34 双曲直线

到这里，有人很自然会问实际的双曲几何是什么样子。也就是说，这个扭曲的地图是反映什么的地图呢？什么东西和圆盘模型的关系，与球面和墨卡托投影的关系是对等的呢？这个问题的确不太好回答。球面几何之所以能够在三维空间中的曲面上实现，这某种意义上是侥幸。如果我们本来是从墨卡托投影开始，从它对距离的奇怪概念开始，并不知道这其实是球面的地图，那么我们要是发现它恰好对应于一个完美对称的曲面，正是将曲面投射到平面的地图，而地图距离只不过是常规意义下容易理解的曲面最短路径长度，那我们一定会既惊讶又高兴。

不幸的是，对于双曲几何来说这种情况却不存在。但有意思的是，这并没有使双曲几何比起球面几何来缺乏实在性。这有点难以理解——至少刚开始有点，但正如我在第二章中强调的，数学概念的实在性更多地与它做什么而不是与它是什么相关。因为我们能够说清双曲圆盘做什么（例如，你要问我，将五边形镶嵌图沿中心五边形的一个顶点旋转30度是什么意思，那我是能够回答你的），所以双曲几何就和其他所有数学概念一样真实。从三维欧氏几何的视角来看，球面几何可能更易于理解，但这并不构成根本的差异。

双曲几何的另一个性质是它满足欧几里得前四条公理。例如，任意两点都可以且仅可以被一条双曲直线段（即与主圆垂直相交的圆弧）相连。尽管看上去好像不能以任意圆心画半径比较长的圆，但这是因为你忘了，越靠近圆盘边缘距离越大。实际上，如果双曲圆几乎要擦上圆盘边缘了，那它的半径（双曲半径）将会非常大。（双曲圆恰好与普通圆外形一样，但它们的圆心却不在我们所料想的位置。参见图35。）

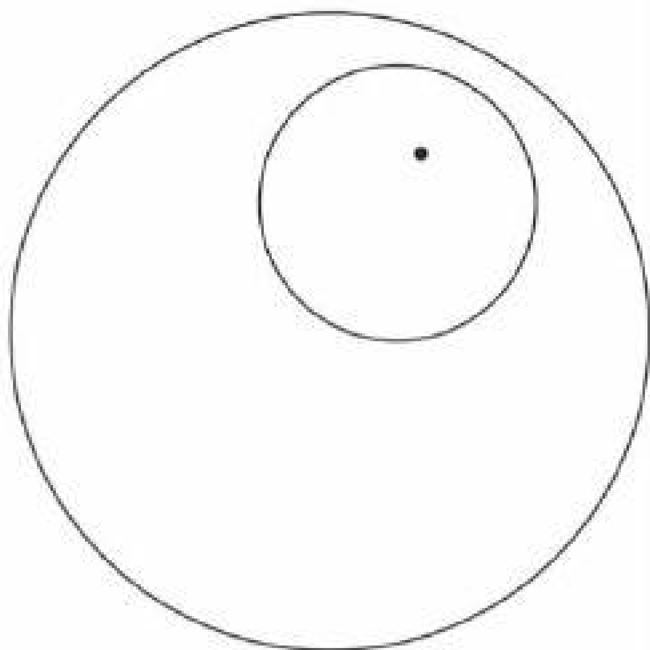


图35 双曲圆及圆心

至于平行公设，如我们所期望的，它对于双曲几何不成立。这可以从图36中看出，在图中我将三条（双曲）直线标为L、M_1和M_2在x点相

交，但它们都不与L相交。因此，经过x有两条（实际上有无穷多）直线不与L相交。这违背了平行公设，平行公设说的是只能有一条直线。换言之，我们在双曲几何中恰好如愿找到了对欧几里得公理新的解释，从而说明了平行公设不是另外四条公理的推论。

当然，在本书中我其实没有证明双曲几何具有我所声称的全部性质。要这样做的话，需要在一般的大学数学课程中花上好几次课的时间，不过至少我还能就如何定义双曲距离说得更确切一些。要说清它，我就必须明确说明，在靠近圆盘边缘处，实际距离比看起来大了多少。答案是，点P处的双曲距离是“常规”距离的$1/d^2$倍，其中d是P到主圆边界的（常规）距离。换一种说法，如果你在双曲圆盘中移动，那么根据双曲距离的概念，你经过P点的速度是你表面速度的$1/d^2$倍。这就意味着，如果你保持恒定的双曲速度，那当你接近圆盘边界时，看起来速度就会越来越慢。

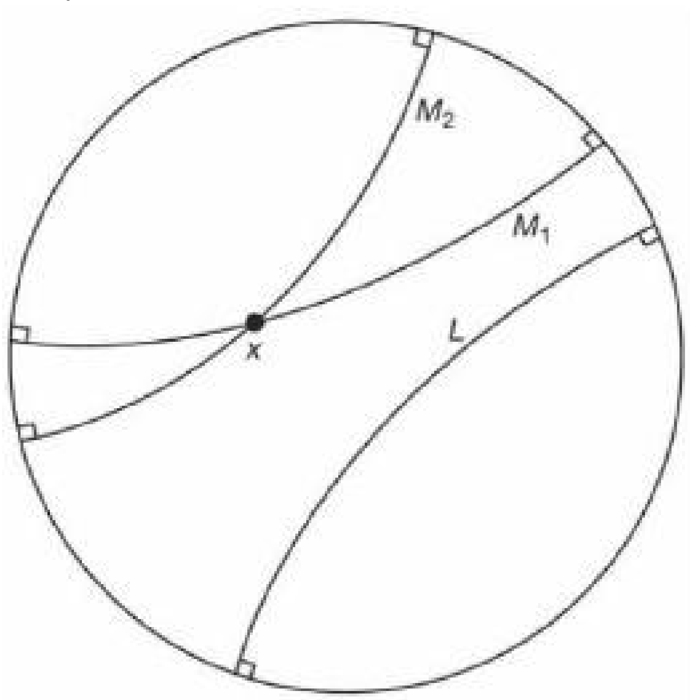


图36 平行公设在双曲平面中不成立

离开双曲几何之前，让我们来看看我前面给出的论证（4）为何不能证明平行线的唯一性。它的想法是这样的：给定直线L和直线外一点x，过x画一条直线M与L不相交，我们可以用几条垂直于L和M的线段将两条线连起来，将L和M之间的空间划分为矩形。这件事看似显然能够做到，但在双曲世界中却是不可能的，因为在其中四边形的内角和总是小于360度。换句话说，在双曲圆盘中，论证所需的矩形是不存在的。

空间何以能够弯曲？

数学中（以及物理学中）听起来最自相矛盾的短语之一就是“弯曲空间”。我们都知道线或面被弯曲是什么意思，但空间本身就是自在之物。即便我们能够在一定程度上对三维空间弯曲的概念赋予意义，与曲面的类比还是揭示出，我们自己不可能观察到空间是否弯曲，除非跳到第四维中去观察。在那里也许我们会发现宇宙是一个四维球体（我在第五章中解释过的概念）的表面，这个球面至少听起来是弯曲的。

当然，这些都是不可能的。因为我们不知道如何能站到宇宙之外——这种想法几乎在措辞上就是矛盾的——我们能够用的证据只能来自于宇宙之内。那么，什么样的证据有可能说服我们空间是弯曲的呢？

和之前一样，如果我们采取抽象方法，这个问题就变得简单些了。不去做艰深的思维体操以试图理解弯曲空间的本性，让我们仅仅遵循扩展数学概念的寻常程序。我们理解“弯曲”一词用在二维表面时的意思。为了把它用在不太熟悉的情形中，即用到三维上，我们必须努力找到弯曲表面易于扩展的性质，就像我们要定义$2^{3/2}$、五维立方体或科赫雪花的维度时所做的那样。因为我们最终要找到的性质应当是能够在空间之内察觉到的，所以我们应当去寻找，怎样在不需要站到弯曲表面之外的情况下就能察觉到它的弯曲性。

比如，确信地球表面是弯曲的一种办法是乘坐航天飞机向地面看，发现它近似球形。但是下面的这个更接近二维情形的试验，同样非常有说服力。从北极点出发一直向正南走大约6200英里，记下你初始的方向。然后向左转，再走相同的距离。然后再向左转，再走一次相同的距离。6200英里大致是北极点到赤道的距离，所以你的旅程会把你从北极点带到赤道，绕赤道走过四分之一，然后再次回到北极点。而且，你回到北极点时的方向应当与你出发的方向夹一直角。于是我们得到，在地球表面上，有一个各角均等于直角的等边三角形。在平坦表面上，等边三角形的内角必须相等且和为180度，所以各角均为60度。因此地球表面不是平坦的。

于是，从曲面内部说明二维曲面弯曲的一种方法，就是找出内角和不为180度的三角形，而且这种办法也是可以在三维中尝试的。这一章我主要关注二维中的欧氏几何、球面几何和双曲几何，但它们都很容易扩展到三维上去。如果我们测量空间三角形的角度，发现它们的和大于180度，这就说明空间更接近于三维版本的球体表面，而不太接近于能用三个笛卡尔坐标描述的那类空间。

如果事实确实如此，那么说空间是正性弯曲的似乎就很合理了。这个空间可以预料到的另一个性质是，沿同一方向出发的直线都会相互靠近并最终相交。还有一个性质，半径为r的圆的周长不是$2\pi r$，要比这个数小一些。

你可能很想指出，我们所知的这个空间并没有这些特性。沿相同方向的直线一直都能保持相同的方向，三角形内角和及圆的周长都如它们所应是的那样。换句话说，即便空间在逻辑上是可能弯曲的，它看起来其实也是平坦的。但是，空间对我们而言看上去平坦，有可能只是因为我们生活在空间特别小的一部分之中。这就像对那些没走过太远的人来说，地球表面看起来也是平的——至少除了大大小小的凹凸外基本平

坦。

换句话说，空间可能只是粗略地平坦。如果我们能够画一个非常大的三角形，或许就会发现它的内角和不是180度。这就是高斯试图做的，可他的三角形却根本不够大。然而，在1919年，史上最著名的科学实验之一显示，弯曲空间的思想不仅是数学家的迷思，更是生活中的事实。根据爱因斯坦四年前发表的广义相对论，空间因引力而弯曲，因而光的行进并不总是沿直线——至少是按欧几里得的意思来理解的直线。这个效应很微弱，很难轻易探测到，但1919年一次日全食提供了很好的机会。这次日全食可以在几内亚湾的普林西比岛观测到。日食发生时，物理学家阿瑟·爱丁顿拍摄下了照片，照片显示靠近太阳的星星并不在通常预期的位置上，这正符合爱因斯坦理论的预测。

尽管人们现在已经接受空间（更精确地讲，是时空）是弯曲的，但也有可能正像地球表面的山峦和谷地一样，我们所观测到的曲率只不过是某个更为庞大、更为对称的形状上的小摄动。天文学中一个重大的未决问题就是去确定宇宙的大尺度形状，即将恒星、黑洞等造成的弯曲熨平后宇宙的形状。它是仍然像大球面一样是弯曲的呢，还是像我们自然而然却很可能错误地想象的那样，是平坦的呢？

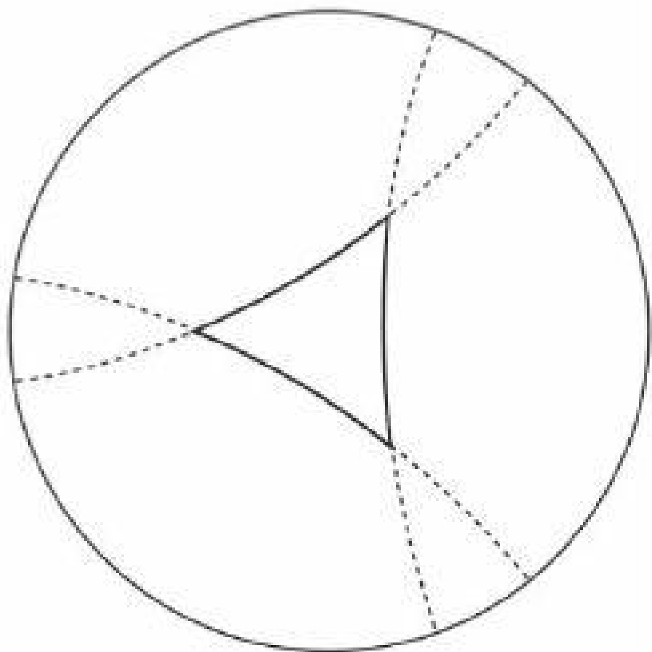


图37 双曲三角形

还有第三种可能性，宇宙是负性弯曲的。容易理解，这是指或多或少与正性弯曲相反。所以，负性弯曲的证据就会是，三角形内角和小于180度，沿相同方向的直线会发散，或者是半径为r的圆的周长大于2πr。这类行为在双曲圆盘上会发生。例如，图37显示了内角和明显小于180度的三角形。要将球面和双曲圆盘扩展到高维的类似情形并不难，而且就大尺度的时空形状而言，双曲几何有可能是比球面几何和欧氏几何更好的模型。

流形

闭曲面是指没有边界的二维形状。球面就是个不错的例子，环面（即铁圈或面包圈的形状的数学称谓）也是。对曲率的讨论表明，曲面虽存在于三维空间之中，但若脱离开三维空间的参照来思考曲面，可能会很有益。如果我们要将闭曲面的概念扩展到高维空间，这样的想法就更加重要了。

不是只有数学家愿意从纯二维的角度来思考曲面。比如，美国的几何结构受到地球弯曲的显著影响，但如果想设计一幅实用的公路地图，则并不需要印在单张弯曲的大纸上。更实际的办法是印成数页的书，每一页上都是这个国家的一小部分。这些局部最好部分重叠，这样一来，如果有城镇位于某一页的边缘，显得不太方便，那它在另一页将不再处于边缘位置。另外，在每一页的边上都要指明，哪一页包含了重叠的区域，重叠以怎样的形式出现。由于地球的弯曲，没有哪一页是绝对精确的，但我们可以在图中画出经线和纬线以指明微小的扭曲。用这种办法，我们就可以用书中平坦的几页纸把美国的几何结构完全包纳了。

原则上，我们可以用类似方法制出精度相当的世界地图册（不过会有很多页几乎全是蓝色的）。因此，球面的数学性质就能以这种方式被一本地图册所涵盖。如果你想要回答有关球面的几何问题，却完全没有能力将球面图像化，只要你手头上有一本地图册，稍加努力就能做得到。图38显示了一个九页的地图册，画的并不是球面，而是环面。要看出它是如何对应于环形曲面的，可以想象一下，将各页粘在一起形成一大页，然后将大页的上下粘合形成圆柱，最后再将圆柱两端粘接。

数学中一大重要分支研究的就是称为流形的对象。流形正是将上述思想扩展到三维或更高维所得的结果。粗略地讲，一个d维流形就是任何一个这样的几何对象，其中任意一点都会被一小块极为类似于d维空间的区域所包围。由于随着维数的增加，流形会变得越来越难以图像化，所以地图册的思想也就相应变得更加有用了。

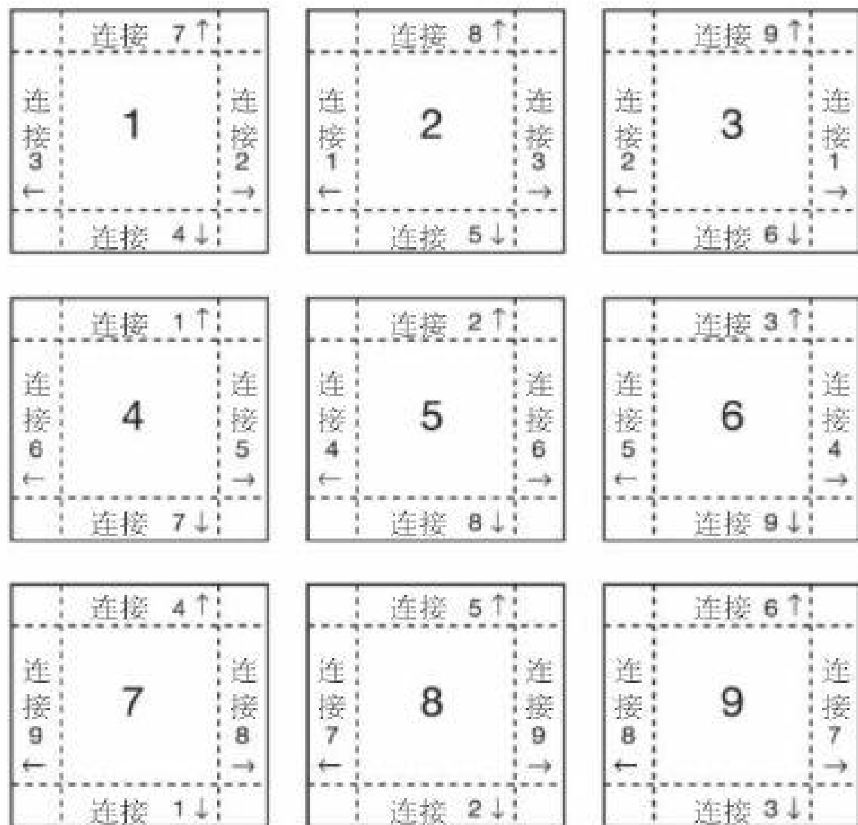


图38 环面地图册

让我们稍加考虑一下，一个三维流形的地图册会是什么样的。其中每一页当然都必须是三维的，而且和公路图的页面一样应当是平坦的。所谓平坦，也就是应当为一块块我们熟悉的欧氏空间。我们可以要求它们都是长方体，但这在数学上并不太重要。这些三维“图页”中的每一页都是流形上一小部分的地图，我们会仔细地标记出各页之间如何重叠。一种典型的标记可能是这样的：在A页某边缘处的（x，y，z）点对应于B页上的（2y，x，z-4）点。

给定了这样的地图册，要如何想象在流形上的移动呢？显然的方式是，设想在某页上移动的一点，一旦该点到达该页边缘时，还会有另外一页表示流形相同的局部，但点在其中不位于边缘，所以我们可以转向那一页。于是，流形的全部几何都可以用地图册来公式化地表达，所以没有必要将流形视作“确实”置身于四维空间中的三维表面。实际上，某些三维流形甚至根本就不可能被放入四维空间中。

一些很自然的问题会随地图册的思想而来。例如，尽管地图册能使我们说明在流形中移动是什么情况，但如果地图册的页数太多，相互重叠的规则特别复杂，我们要如何从中得到对流形基本“形状”的一些感觉呢？我们怎样才能辨别两本不同的地图册其实都是关于相同流形的？特别是，有没有某种简易的办法，让我们一看到三维地图册，就能辨别它表示的是不是四维球体的三维表面？最后一个问题的一种精确表述称为

庞加莱猜想，尚未解决；解答这一问题将能够得到100万美元的奖金（由克雷数学研究所提供）[1]。

【注释】

[1] 庞加莱猜想现已由俄罗斯数学家佩雷尔曼最终解决。

第七章
估计与近似

大多数人认为数学是一门纯净、精确的学科。我们经过在中小学的学习，料想数学问题如果可以被简洁地陈述，大概就能得到简练地回答，通常是一个简单的公式。而继续学习大学阶段数学的人，尤其是那些专门研究数学的人，很快就发现这样的想法实在是大错特错。对于很多问题来说，如果有人能够找到解答的精确公式，那简直完全出人意料，如同奇迹一般。多数情况下，我们不得不满足于大致的估计。在你对此感到习以为常之前，这些估计总是看似很丑陋，难以令人满意。然而，品尝一下其中的滋味也是值得的，否则你就会错过数学中很多最伟大的定理以及最有趣的未解决问题。

无法用简单公式表达的简单序列

记a_1，a_2，a_3，…为一列实数，根据如下规则产生。第一个数a_1等于1，后面各个数都是前一个数加上其平方根。也就是，对任意n，我们让$a_{n+1}=a_n+\sqrt{a_n}$。基于这个简单的规则，显然可以提出问题：a_n是什么数？

为了对这个问题有些感性认识，让我们对较小的几个n来算出a_n。我们有$a_2=1+\sqrt{1}=2$。接下来：等等等等。注意，等号右端的表达式看起来无法简化，而且每个新的表达式都是前一个长度的两倍。从这个观察中，很容易得出，数字a_{12}的表达式中将会出现1024个2，它们中的大多数都深深地位于根号丛中。这样的表达式并不会使我们对数字a_{12}有多深的了解。

$$a_3=a_2+\sqrt{a_2}=2+\sqrt{2}$$

$$a_4=a_3+\sqrt{a_3}=2+\sqrt{2}+\sqrt{2+\sqrt{2}}$$

$$a_5=a_4+\sqrt{a_4}=2+\sqrt{2}+\sqrt{2+\sqrt{2}}+\sqrt{2+\sqrt{2}+\sqrt{2+\sqrt{2}}}$$

那么我们因此就应当放弃去理解这个序列吗？不。尽管当n比较大时，要考虑a_n的精确值似乎没有好办法，但这并未排除得到较准确估计的可能性。实际上，一个好的估计最终可能反而会更有用处。我上面写出了a_5的精确表达式，但和a_5约为$7\frac{1}{2}$这条信息比起来，这个表达式让你更好地理解a_5了吗？

所以，让我们不要再问a_n是什么，转而去问a_n大约是多大。也就是，让我们寻找能给出a_n较优近似值的简单公式。这样的公式是存在的：a_n大约为$n^2/4$。要严格地证明需要一点技巧，但要看出为什么这个估计是合理的，我们要注意：

$$\frac{(n+1)^2}{4}=\frac{n^2+2n+1}{4}=\frac{n^2}{4}+\frac{n}{2}+\frac{1}{4}=\frac{n^2}{4}+\sqrt{\frac{n^2}{4}}+\frac{1}{4}$$

也就是说，如果$b_n=n^2/4$，那么$b_{n+1}=b_n+\sqrt{b_n}+1/4$。要是没有其中的“＋1/4”的话，这个式子就会告诉我们b_n正和a_n的生成方法一样。而当n较大时，加上1/4“只不过是个小扰动”（这就是我略去证明的部分），所以b_n可以看作近似正确地生成，由此推出b_n即$n^2/4$给出了a_n较好的近似，正如我之前所断言的。

近似的方法

作类似于刚才的关于近似的论断时，明确什么才算是较好的近似很重要，因为标准会随着情景的不同而变化。如果想用一条能够简单定义的序列b_1，b_2，b_3，…来近似一条稳定增大的序列a_1，a_2，a_3，…，那我们所能期待的最优近似——实则很少能够达到，就是每一对a_n和b_n的差距都小于一定值——诸如1000。那么随着a_n和b_n增大，它们的比值会非常接近于1。例如，假设某时有$a_n=240859734863249875882 8$，以及$b_n=2408597348632498759734$。那么$b_n-a_n=906$，这个差虽然是个大数，但与$a_n$和$b_n$比起来就微不足道了。如果$b_n$是在这种意义上对$a_n$的近似，我们可以说$a_n$和$b_n$是“相差常数以内的相等”。

另一类较优的近似是，随着n的增大，a_n和b_n的比值变得非常接近1。当a_n和b_n相差常数以内相等时，这种情况是成立的，但它在其他一些情况下也会成立。例如，如果$a_n=n^2$而$b_n=n^2+3n$，则比值b_n a_n为$1+3n$，在n较大时很接近1，尽管a_n和b_n的差3n会很大。

即便这种情况常常也是难以企及的，如果能找到更弱的近似方法我们就会很高兴了。常见的一种方法是，如果a_n和b_n“相差常数倍以内相等”，就视之为近似相等。它的意思是，a_n/b_n和b_n/a_n都不会超过某个固定的数——诸如1000也是可能的，但越小越好。换言之，此时不是a_n和b_n的差，而是它们的比值保持在某个限度内。

要说两个相差1000倍的数大致相同，这或许看似很不合情理。但这是因为我们总习惯于处理较小的数。自然没有人会将17和13 895看作大致相同，但宽泛地讲，要说下面这样两个数规模相类则并不显得很荒唐：

290475629408976156238954534598760889079687234751434875777546 8和

360982345987209761234987098249863408762345779678458734598716 6464

尽管第二个数比第一个数大1000倍以上，但它们的数位大体相同——都在60到65位之间。既然没有其他有趣的性质，这一点很可能就是我们所关注的全部。

如果连这样程度的近似都是奢求，那也常常值得去找出两条参考序列b_1，b_2，b_3，…和c_1，c_2，c_3，…，并证明b_n总小于a_n，而c_n总大于a_n。那么我们可以说b_n是a_n的“下界”，c_n是a_n的“上界”。比如，一位数学家想估计某个量a_n，他可能会说：“我不知道a_n的值是多少，甚至连近似值也找不到，但我能够证明它至少是$n^3/2$，但不大于n^3。”如果这个问题相当困难，那这样的一个定理就可能成为重大进展。

关于对数、平方根等你只需要知道这些

估计和近似遍布于数学之中，而这个领域以外的人却不太了解这一点，部分原因在于，为了谈论近似，要使用类似于“大约和logn的速度一样快”，或者“限制在相差一个常数内为根号t”这样的语言，而这对多数人来说意义不大。幸而，对于较大数的对数或平方根，如果有人只关心近似值，便很容易理解，这一类语言也不难懂。

如果你取两个大的正整数m和n，想快速得出乘积mn的估值，那你应该怎么做？较好的办法是，一开始先数一下m和n各自的位数，如果m有h位，n有k位，那么m位于10^{h-1}和10^{h}之间，n位于10^{k-1}和10^{k}之间，所以mn位于10^{h+k-2}和10^{h+k}之间。于是，仅仅数一数m和n的位数就可以“在100倍以内”决定mn，即mn必落在10^{h+k-2}和10^{h+k}这两个数之间，且10^{h+k}仅比10^{h+k-2}大100倍。如果你折中一下，取10^{h+k-1}作为估计值，那么它与mn相差的因子至多为10。

换句话说，如果你只在“相差常数倍以内”的程度上关心某个数，那么乘法立刻就变简单了：取m和n两个数，数数它们合起来的位数，减去1（如果你在乎这一点），写下一个有这么多位的数。例如，1 293 875（7位）乘以20 986 759 777（11位）大致接近10 000 000 000 000 000（17位）。如果你想做得更仔细一点，那么可以用上第一个数的首位1及第二个数的首位2，也就意味着20 000 000 000 000 000是个更优的估计值，不过在很多场合下这样的精度是不必要的。

既然近似乘法这么简单，近似平方也就很简单了——只要把原来的数替换为一个两倍位数的新数。由此得出，把n的位数减半可以求出n平方根的近似值。类似地，将位数除以3可以求出立方根的近似值。更一般地，如果n是个大整数，t是任一正数，那么n^{t}的位数大约为n的位数乘以t。

关于对数呢？从近似的观点看它们其实极为简单：一个数的对数基本上就是它所包含的位数。例如，34587和492 348 797 548 735的对数分别约为5和15。

实际上，数一个数的位数所近似的是它以10为底的对数，即在便携计算器上按LOG键得到的数。通常，数学家谈论对数指的都是所谓的“自然”对数，也就是以e为底的对数。尽管e这个数的确很自然也很重要，但我们在此只要了解：一个数的自然对数，即在计算器上按LN键得到的数，大体上是它的位数乘以2.3上下。于是2 305 799 985 748的自然对数约为：13×2.3＝29.9。（如果你了解对数，你会知道真正应当乘上的数是$\log_{e}10$。）

这一过程也可以反过来进行。假设你有一个数t且知道它是另一个数n的自然对数。那么数n称作t的指数幂，记作e^{t}。那么n大致是多少呢？刚才为了得到t，我们数n的位数并乘以2.3。所以n的位数必定约为t/2.3。这就决定了n的大小，至少是近似值。

我上面给出的近似值定义最主要的用途是作比较。例如，现在我们能够明显看出，对大数n而言，其对数要远小于立方根：比如，如果n有75位，它的立方根将会很大——大约有25位数，但它的自然对数则大约仅为75×2.3＝172.5。类似地，数m的指数幂会远大于它的乘方，如m^{10}。例如，若m有50位，那么m^{10}大约有500位，但e^{m}大约有m 2.3

位，远大于500。

对数字n＝941 192施行不同的运算，得到的近似结果如下表所示。我没有包含进e^n，因为如果要这么做，那我就得给这本书换个名字了。

n	941192
n^2	885842380864
$\sqrt{n}$	970.15
$\sqrt[3]{n}$	98
$\log_e n$	13.755
$\log_{10} n$	5.974

素数定理

素数是大于1且不能被其他整数——1和自身显然除外——整除的整数。小于150的所有素数有2，3，5，7，11，13，17，19，23，29，31，37，41，43，47，53，59，61，67，71，73，79，83，89，97，101，103，107，109，113，127，131，137，139和149。小于150的其他数都可以做因子分解：例如91＝7×13。（你可能会担心，将1排除在素数的定义之外似乎不合情理。这并不表达数的某种深层事实，只是碰巧成为一个有用的惯例，采纳这样的定义能使任意数被分解为素数的方式仅有一种。）

自从古希腊时期以来，素数就一直困扰着数学家们，因为它们表面上多多少少是随机分布的，但又并非全然随机。从没有人找出一种简单的规则，能够告诉我们第n个素数是多少（当然可以下番工夫列出前n个素数，但这算不上是简单的规则，而且当n很大时会很不切实际），几乎也不太可能会有这么一种规则。但另一方面，仅仅检查前35个素数，也能向我们透露一些有意思的特征。如果你算出相邻两个素数的差，那么你会得到下面这个新的序列：1，2，2，4，2，4，2，4，6，2，6，4，2，4，6，6，2，6，4，2，6，4，6，8，4，2，4，2，4，14，4，6，2，10。（即1＝3-2，2＝5-3，2＝7-5，4＝11-7等等。）这列数仍显得有些不规则，但这些数有一个趋势，大致可以看出它们逐渐在增大。显然它们不是稳定地增大的，但最开始的几个数都不大于4，而10和14这样的数直到后来才会出现。

如果你写出前1000个素数，那么相邻素数间距增大的趋势就会变得更加明显。也就是说，和小素数比起来，大素数的出现越来越稀疏。我们正可以预料到这一点，因为大数不成为素数的方式更多了。例如，可能会有人猜测10 001是素数，尤其因为它无法被2，3，5，7，11，13，17以及19整除——但实际上它等于73×137。

有自尊心的数学家绝不会满足于仅仅观察到（甚至没有得到严格地证明）大素数比小素数稀少。他一定想要知道，它们稀少到何种具体程度。如果你在1 000 001和1 010 000之间随机取一数，那么这个数有多大的机会是素数？换言之，1 000 000附近的素数“密度”是多大？它是极其小还是仅仅比较小？

没有接触过大学数学的人很少会提出这样的问题，其原因在于，他们没掌握将问题公式化表达并进一步思考所需的语言。不过，若是这章到现在为止你都看懂了，那么你就能够欣赏到数学中最伟大的成就之一：素数定理。定理陈述的是，在数n附近的素数密度约为$1/\log_e n$，即1除以n的自然对数。

现在再来考虑1 000 001和1 010 000之间取随机数为素数的机会大小。这个区间内的数都大体等于100万。按照素数定理，密度因而约为1除以100万的自然对数。它以10为底的对数是6（在这个例子中，数位数会得到7，但既然我们知道精确答案，不妨使用精确值），所以自然对数约为6×2.3，即13.8。因此，1 000 001和1 010 000之间平均每14个数中有1个是素数，即略多于7 。相比之下，小于100的素数共有24个，即几乎占所有数的四分之一，这就说明了随着数的增大，素数密度是如何减小的。

既然素数分布有零零散散、颇似随机的性质，而我们却能证明其如此多的特点，这足以令人十分惊讶。有意思的是，关于素数的定理通常都是通过利用这种看似随机的性质得到证明的。例如，维诺格拉多夫在1937年证明的一个著名定理认为，任意充分大的奇数都可以分解为三个素数之和。我无法在本书中解释他是怎样证明的，但他绝对没有找出将奇数表达为三素数之和的方法。这样的思路几乎注定会失败，因为即使是生成这些素数本身也非常困难。基于哈代和利特伍德之前的工作，他大体按照下述办法来论证。如果你能够按照和素数分布同样的密度来真正随机地选取一些数，那么概率论的某种初步理论就能够表明，你几乎一定能够将所有充分大的数表示为你所取的这些数中的三个之和。实际上，你能够以多种不同方式进行这一分解。因为素数是类似于随机的（证明中较难的部分就是要说明，“类似于随机”是什么意思，再加以严格证明），它们的行为就相仿于随机选取的序列，所以所有充分大的数都是三素数之和——同样也以多种方式。为了解释这种现象，这里我们以35为例，列出它分解为三素数之和的所有方式：

$$35=2+2+31=3+3+29=3+13+19=5+7+23$$
$$=5+11+19=5+13+17=7+11+17=11+11+13$$

关于素数的很多研究都具有此类特点。你首先对素数设计一种概率模型，即假装告诉自己，它们是根据某种随机过程挑选出来的。接下来，在假设素数的确是随机产生的情况下，求证有哪些论断是正确的。这样可以使你猜测出很多问题的答案。最后，你努力表明，这个模型足够现实，能够保证你的猜测近似准确。要注意的是，如果强迫在论证中的每一步都给出精确答案，那这个思路就是不可能的。

很有意思，概率模型不仅仅是物理现象的模型，还能成为另一数学分支的模型。尽管素数的真实分布是严格确定下来的，可某种程度上它们看起来也像是实验数据。一旦这样看待它们，我们就很想去设计对应的简化模型，来预测特定概率论问题的答案是什么样的。这种模型有时的确曾使人们得到对素数本身的有效证明。

这种风格的论证取得了某些显著的成功，但它依然没能解决许多著名问题。例如，哥德巴赫猜想断言，任意大于4的偶数都可以表示为两个奇素数之和。这个猜想看起来比维诺格拉多夫所解答的三素数猜想要难得多。此外还有孪生素数猜想[1]，它声称有无穷对相距为2的素数，诸如17与19，137与139。表达这个问题的另一种方式是，如果你写出相邻素数之差，就像我之前做过的那样，那么2永远都会出现（尽管越来越稀少）。

数学中最著名的未决问题大概要数黎曼假设了。这个假设有若干种等价的表达形式，其中一种涉及素数定理给出的估计的精度。我之前说过，素数定理告诉我们在某数附近素数的近似密度。根据这个信息，我们可以计算出，不大于n的素数大约有多少。但这个“大约”有多“大约”？如果p（n）是不大于n的素数个数的真实值，q（n）是根据素数定理得到的估计值，那么黎曼猜想断言，p（n）和q（n）的差不会比$\sqrt{n}$大太多。如果能够证明这样的精度确实成立，那么它将会有很多的应用，但迄今所得到的结果比这要弱得多。

排序算法

数学另有一个分支领域，理论计算机科学，其中满是粗略的估计。如果有人要写一个计算机程序完成特定任务，那么程序当然运行得越快越好。理论计算机科学家提出了这样的问题：我们所能期望的最快速度是多快？

要精确地回答这个问题几乎总是不现实的，所以我们转而证明诸如“当输入规模为n时，这种算法的运行步数约为n^2”这样的论断。由此出发我们可以总结出，一台普通的个人电脑有可能处理的输入规模（简言之，即你想要分析的信息量）为1000，而不是1 000 000。于是，这样的估计就有了实际的重要性。

计算机能够完成的一项非常有用的任务被称为排序，即将为数很多的对象按照给定标准排列顺序。举例来说，想象一下你想要根据喜好程度来排列一系列对象（不一定非要是无生命的，也有可能是一项工作的应聘者之类）。设想，你无法对每个对象赋予一数值来表示有多喜欢它，但是若给定任意两个对象，你总能确定更喜欢哪一个。再设想，你的偏好是前后一致的，意思是指，永远不可能出现这样的情况：喜欢A胜于B，喜欢B胜于C，又喜欢C胜于A。如果你不想为这个任务花太多时间，那么将比较次数最小化就会很有意义。

当对象很少时，很容易看出要怎样做。比如有两个对象，那你至少要作一次比较，一旦比较了你也就知道了它们的排列顺序。若有三个对象A、B、C，那么一次比较是不够的，但总得从某个比较开始，从哪个开始都无妨。为了论证，假设你先比较了A和B，更喜欢A。这时你必须用A和B中的一个和C进行比较。如果你用A和C来比较，并且更喜欢C，那么你就知道了你的偏好顺序是C，A，B；但如果你发现更喜欢A，那么关于B和C，你就只知道它们都没有A讨你喜欢。所以这时需要进行第三次比较，以便能将B和C加入排序。综上所述，三次比较总是充分的，而且有时也是必要的。

有四个对象A、B、C、D时情况如何呢？分析起来就更加困难了。你可能仍从比较A和B开始。但一旦比完这一对，接下来的比较就会有两种本质不同的可能性。一种是用A或B来比较C，另一种是比较C和D，哪种主意更好并不明显。

设想你要比较B和C。如果幸运，你就能够对A、B、C进行排序。假设排序就是A，B，C。接下来还需要考虑D适合什么位置。最好的办法是先比D和B。在这之后，你就只需再比较一次，即比较D和A（如果D比B好）或比较D和C（如果B比D好）。以上一共比较了四次——两次来确定A、B、C的顺序，两次找出D的位置。

我们还没有分析完这个问题，因为在A、B、C的比较中你可能并不走运。也许比完前两次以后，你只知道A和C都比B好。那么面临着另一个困境：是比较A和C更优呢，还是用D来与A、B和C中的一个比较更优呢？考虑完各种情况及各种子情况，你还需要再观察，如果第二次比较的是C和D，情况会如何。

分析变得烦琐起来，但终究还是能够完成的。它表明，五次比较总是足够的，有时必须要比较五次之多，且第二次比较应当在C和D间进行。

这种论证的麻烦在于，要考虑的情况数量会急剧增加。譬如当对象有100个时，需要比较的确切次数是不可能算出来的——几乎可以确定这个数永远无从得知。（我犹记得，当我第一次听一位数学家宣称某个特定量的确切值永远无法算出，我当时有多么震惊。但现在我已经习以为常了，这是常有的事，并不是特殊情况。当时的那个数是拉姆齐数R（5，5），即为了保证n个人中必有五人全部两两认识，同时必有五人全部两两陌生，那么n至少是多大。）作为替代，我们要去寻找上界和下界。对于这个问题来讲，上界c_n的意思是，对n个物体进行排序的某种过程至多进行c_n次比较，下界b_n的意思是：无论你有多聪明，有时候也必须要做上b_n次比较才行。这个例子中，已知的最优上界和下界正相差一个乘数因子：相差常数倍情况下，对n个对象进行排序所需的比较次数为nlogn。

要理解这为什么有意思，你不妨自己构思一种排序过程。一种明显的方法是，先找出排序最靠前的对象放到一边，接下来再重复进行。为了找到最好的对象，先比较头两个，再将胜出者与第三个比，再将胜出者与第四个比，依此类推。这种办法下，需要比较n-1次才能找到最好的，再找出次好的需要比较n-2次，等等等等，于是总的比较次数为（n-1）+（n-2）+（n-3）+…+1，计算得约为$n^2/2$。

尽管这种方法很自然，但如果你采用它，最终排列完时就比较了所有对象中的每一对，所以它其实是最低效的方法（虽然它的确具有易编程的优点）。当n很大时，nlogn是对$n^2/2$很显著的改进，因为logn比n/2要小得多。

下述方法被称为“快速排序”，它并不能保证一定就会更快，但通常情况它都会快很多。它是如下递归（即利用自身）定义的。首先选择任一对象，比如x，再将其他对象分成两堆，一堆全都优于x，另一堆全都劣于x。这需要比较n-1次。接下来你只需分别对两堆进行排序——再次利用快速排序。也就是，对每一堆来讲，选出一个对象，将剩余对象分成两小堆，依此类推。一般来说，除非你运气很差，否则分出的两堆中对象数量是差不多的。于是可以表明，所需比较次数大致为nlogn。换句话说，这种方法通常都能以你期待的效率运作，相差常数倍以内。

【注释】

[1] 2013年，华人数学家张益唐证明了弱形式的孪生素数猜想。

第八章

常见问题

1. 数学家在30岁以后就不比当年了，这是真的吗？

这种传说影响颇为广泛，正由于人们误解了数学能力的本质，才使得它很有吸引力。人们总喜欢把数学家看作极具天资的人，并认为天资这种东西有些人生来就有，其他人则绝难获得。

其实，年龄与数学成果间的关系对不同人来说差别很大。的确有一部分数学家在20来岁的时候做出了他们最杰出的工作，但绝大多数人都认为，他们的知识水平和专业素质终其一生都在稳健地提高，在许多年里，这种专业水平的增长都能够弥补“原生”脑力的任何衰退（如果确实有“原生”脑力这回事的话）。确实数学家在年逾40岁之后就少有重要的突破性进展了，但这也很有可能是社会学方面的原因。到了40岁时，如果有人还有能力做出突破性的工作，那么他极有可能早已因之前的工作闻名遐迩，因而有可能也不像未成名的年轻数学家那样具有奋斗精神。不过还是有很多反例的，有很多数学家在退休之后热情不减，还继续在数学领域工作。

一般来讲，人们通常所想象的数学家的形象——可能很聪明，但有点古怪，穿着邋遢，毫无性欲，比较孤僻——的确不是一种讨喜的形象。有一部分数学家在一定程度上的确符合这种形象，但如果你认为不这样就做不好数学，这种想法可就太蠢了。实际上，如果所有其他条件都相同的话，可能你还要比这些怪数学家更胜一筹。一开始学习数学的所有学生中，最后成为专职研究人员的比例极小，更多的人在早期阶段便离开了数学，比如失去兴趣、没有申请到读博机会，或者得到了博士学位但没有获得教职。在我看来（实际上不仅只有我这么想），对最终通过了这层层考验的人来说，那些“怪人”所占的比例比占一开始学习数学的学生的比例要小。

对数学家这样的负面刻画可能杀伤力很大，吓走许多本来可能喜欢并且擅长这一领域的人，但是“天才”这个词则更加恶毒，杀伤力更大。这里有一个现成的对“天才”的大致定义：对于别人必须经过多年实践都未必能够掌握的事情，天才就是那些在少年时期就能够轻易做好这些事的人。天才的成就有着魔法般的特质，就好像他们的大脑并不只是比我们更有效率，而是运转方式完全不同。剑桥大学每年都会有一两个数学系本科生，他们经常在数分钟之内就能解决的问题，大多数人——包括应该能够教他们的人——往往需要花上几个小时以上。遇到这种人的时候，我们只能退避三舍、顶礼膜拜。

然而，这些超乎寻常的人并不总是最成功的数学研究者。如果你想要解决某个问题，而之前尝试过的数学家都以失败告终，那么你需要具备种种素质，在这其中天赋（如我所定义的那样）既不是必要的也不是充分的。我们可以通过一个极端一点的例子来说明这一点。安德鲁·怀尔斯（在刚到40岁的时候）证明了费马大定理（即对任意正整数x，y，z及大于2的正整数n，x^n+y^n不可能等于z^n），解决了世界上最著名的数学难题。毫无疑问他很聪明，但他并不是我所说的天才。

你可能会问，如果没有某种神秘的超常脑力，他还可能完成这一切吗？回答是，尽管他的成就非常卓著，但也没有卓越到无法解释的程度。我并不了解究竟是什么因素促使他成功的，但他肯定需要非凡的勇气、坚定和耐心，对他人完成的艰难工作的广泛了解，在正确时间专攻正确领域的运气，以及杰出的战略性眼光。

上面所说的最后一条素质，从根本上要比惊人的大脑运转速度更加重要。数学中绝大多数影响深远的贡献是由“乌龟”们而不是“兔子”们做出的。随着数学家的成长，他们都会逐渐学会这个行当里的各种把戏，部分来自于其他数学家的工作，部分来自于自己对这个问题长时间的思考。是否能将他们的专长用于解决极其困难的问题，则在很大程度上决定于细致的规划：选取一些可能会结出丰硕成果的问题，知道什么时候应该放弃一条思路（相当困难的判断），能够先勾勒出论证问题的大框架继而再时不时地向里面填充细节。这就需要对数学有相当成熟的把握，这绝不与天赋相矛盾，但也并不总是会伴随着天赋。

2. 为什么女性数学家很少见？

真想回避这个问题，因为回答这个问题很容易冒犯别人。但是，在全世界各地的数学系所中，即便是在今日，女性所占比例仍然很小；这是一个值得注意的现象，也是数学生活中的一个重要事实，我被迫感到不得不说点什么，尽管我所要说的也无非是对此感到不解和遗憾。

值得强调的一点是，数学家中女性较少只不过是一种统计现象：确实有十分优秀的女性数学家，与男性同行一样，她们表现优秀的方式也多种多样，有时也包括拥有天赋。没有任何迹象表明，女性在数学中所能达到的成就会有上限。我们有时会读到，在特定的智力测试中——比如说视觉空间能力，男性表现得更优秀，有人认为这解释了他们主导着数学领域的原因。然而，这样的论据不足以令人信服，因为视觉空间能力能够通过练习来增强，而且尽管它有时对数学家有帮助，却并非不可或缺。

更可信的一种理由是社会方面的因素：当男孩子为数学能力感到骄傲时，可以想象某个女孩子可能会为自己擅长这项不那么女性化的事务而感到窘迫。而且，有数学天赋的女孩子所能够效仿的榜样很少，她们只能靠自我保持、自我强化。一项社会因素可能会在之后发挥更大的作用：比起其他学科来，数学需要一个人更加专注，这虽然不是不可能，但也很难与女性的母亲身份相结合。小说家坎迪亚·麦克威廉曾经说，她的每个孩子都使她少写了两本书，不过在几年未动笔之后，她至少还能够重新写小说。但如果你几年没有做数学，你就失去了数学的习惯，很难再重拾了。

有人认为，女性数学家发展起自己事业的时间往往晚于男性同行，而数学家的职业结构倾向于回报早期成就，这就使得女性处于一种不利的地位。最杰出的一些女性数学家的人生故事支持了这种说法。不过她们发展自己职业生涯较晚的原因，基本上都是上面所说的社会原因，而且也有许多这方面的例外。

不过，这些解释看起来都不够充分。我在此不再深入探讨了。我还能做的就是告诉大家，关于这方面已经出了几本书（参见“延伸阅读”）。最后再加上一点评论：这样的情况是在不断进步的。数学家中女性所占的比例近年来在稳步提高，而且随着社会大环境的不断改变，这样的现象一定还会持续下去。

3. 数学与音乐息息相通吗？

尽管有很多数学家完全不了解音乐，也很少有音乐家对数学感兴趣，但一直有一种民间观念认为这两个领域是相关联的。其结果就是，当我们听说某位数学家钢琴弹得很好，或者爱好作曲，或者喜欢听巴赫，没有人会对此感到惊奇。

有很多奇闻轶事在讲，各种艺术形式中，数学家为音乐所吸引的最多。也有一些研究声称已经表明，受过音乐教育的儿童在科学领域中表现得更优秀。我们不难猜出为什么会这样。尽管在所有艺术形式中抽象都很重要，但音乐在其中最具有代表性，可以说是最明显的抽象艺术：听音乐所获得的愉悦感，大部分来自于对不具有内在含义的纯粹形式的直接——即使不是完全自觉的——欣赏。

不幸的是，这些传说中的证据很少得到严格的科学支持。关于这种说法，就连应该提出哪些疑问都不好说。如果我们收集到统计数据显著地说明，在相近的社会背景及教育背景下，数学家与其他人相比，弹钢琴的百分比更高，那我们能够从中了解到什么呢？我自己猜测，的确会得到这样的数据。但如果提出一种可经实验验证的理论来说明这其中的关联，会有趣得多。就统计证据而言，如果能够更加详尽明确，也会更有价值。数学和音乐都是内容很广泛的领域，某些人很有可能只对领域中的某一部分有热情，而对其他部分毫无兴趣。数学和音乐趣味之间是否会有微妙的联系？如果有，那将会比这两个领域间整体的粗略相关性更具信息含量。

4. 为什么有那么多人旗帜鲜明地厌恶数学？

我们不常听到别人说他们从来不喜欢生物学，或者英国文学。毫无疑问，并不是所有人都会对这些学科感到兴奋，但是，那些没有热情的人往往完全理解那些有热情的人。相反，数学，以及其他内容高度数学化的学科，诸如物理，似乎不仅仅使人提不起兴趣，而且能激起反感。究竟是什么原因使他们一旦能够抛弃数学时就立刻抛弃，并且一生都对数学心有余悸？

很可能并不是因为数学很无聊，而是数学课的经历很乏味。这一点更容易理解。因为数学总是持续在自身的基础上构建，所以学习时的步步跟进就显得很重要。比方说，如果你不太擅长两位数的乘法，那你很可能就不会对分配律（第二章中讨论过）有良好的直觉。没有这种直觉，你可能就会在计算打开括号（$x+2$）（$x+3$）时感到不适应，于是你接下来就不能很好地理解二次方程，因而也无法理解为什么黄金分割比是 $\frac{1+\sqrt{5}}{2}$。

类似这样的环环相扣还有很多，但是，学习数学时的步步跟进不仅仅是保持技术熟练度而已。数学中常常会引入重要的新思想，新思想会比旧思想更加复杂，每一个新思想的引入都有可能把我们甩在后面。一个很明显的例子就是用字母表示数，很多人对此糊里糊涂，但对某个层次以上的数学来讲这是基础性的。还有其他类似的例子，比如负数、三角函数、指数、对数以及初步的微积分。没有作好准备来进行必要的概念飞跃的人，一旦遇到这些新思想时，就会对其后建立在新思想基础上的一切数学感到并不牢靠。久而久之，他们就会习惯于对数学老师所说的东西仅仅一知半解，日后再错过几次飞跃，恐怕连一知半解也做不到了。同时他们又看到班上其他同学能够轻而易举地跟上课程。因此就不难理解，为什么对许多人来讲数学课成为了一种煎熬。

情况一定是这样的吗？有没有人天生注定就会在学校里厌恶数学，还是说，有可能找出一种不同的数学教学方法，使得排斥数学的人能够大大减少？我相信，小孩子如果在早期接受到热情的好老师一对一教学，长大之后就会喜欢上数学。当然，这并不能直接成为一种可行的教育政策，不过至少告诉我们，数学的教育方法可能有改进空间。

从我在本书中所强调的思想出发，我可以给出一条建议。在上面，我间接地将技术的熟练度与对较难概念的理解作了一番比较，但实际情况似乎是，凡是擅长其中一个方面的必然两个方面都擅长。况且，如果说理解数学对象，大体上就是要学习数学对象所遵从的规则，而非把握其本质，那么我们完全可以预期：技术的熟练度与数学理解力之间并不像我们想象得那样泾渭分明。

这又会对课堂实践产生什么影响呢？我并不赞成革命性的改进——数学教育已经深受其累，我所赞同的是小幅度的改变，有所侧重的小幅变化将会是有益的。比方说，一个小学生犯了个常见错误，觉得$x^{a+b}=x^a+x^b$。强调表达式x^a内在含义的老师会指出，x^{a+b}的含义是$a+b$个x相乘，显然与a个x相乘再乘以b个x相乘的结果相等。不幸的是，很多孩子觉得这样的论证过于复杂、难以领会，何况一旦a和b不是正整

数，这样的说法就无效了。

如果使用更抽象的办法，那么这些孩子可能会从中获益。正如我在第二章中所指出的，关于指数我们需要了解的一切，都能从几条很简单的规则中推导出来，其中最重要的一条就是$x^{a+b}=x^ax^b$。如果这条规则得到了强调，那么上面的这种错误可能出现的机会就减少了，一旦出现了也很容易纠正：我们只需要提醒犯错的人没有使用正确的规则就行了。当然，熟悉x^3等于x乘以x乘以x这样的基本事实也很重要，但这样的事实可以当作规则的推论出现，而不是当作规则的论据。

我并不是想说，我们应该向孩子们解释什么是抽象方法，我只是想指，教师们应当对抽象方法的隐含意义有所认识。这些隐含意义中最主要的一个就是，即使并不能确切地了解数学概念的含义，我们也很有可能学会正确地使用它们。这听起来似乎是个坏主意，但是用法总是容易教，而对意义的深层理解——倘若在用途之上的确有某种意义的话——常常会自然而然地随之而来。

5. 数学家在工作中使用计算机吗？

简单地说，大多数数学家并不用，或者说，即使用也不会在数学工作中占据基础性的地位。当然，正如所有其他人一样，我们也认为，在文字处理和交流沟通方面计算机是不可或缺的，互联网也在日益发挥着重要的作用。在某些数学领域中，人们必须常规性地做冗长烦琐的基本计算，有一些很不错的符号计算程序能够完成这些工作。

所以，计算机能够成为非常有用的节省时间的机器，某些时候，它们甚至能使数学家发现单靠自己无法发现的结果。不过，计算机能够提供的帮助还是很有限的。如果你所研究的问题——更多情况下是子问题——正好是那一小部分能够通过长时间重复性工作完成的，那当然不成问题。可如果你的研究受阻，需要一个聪明的想法，那么就当前的技术状况而言，计算机什么忙也帮不上。实际上，大多数数学家都会说，他们最重要的工具还是纸和笔。

我的观点属于少数派，我认为这种情况只是暂时的，在未来的一百年左右的时间里，计算机将渐渐能够为数学家做越来越多的工作——没准会从帮我们做一些简单的练习开始，或者防止我们在证明错误的引理上浪费一个星期，以免到时才发现一种著名的构造就给出了反例（我说的正是自己常常经历的），直到最终完全取代我们。大多数的数学家则远比我悲观（或者说这才应该属于乐观？），认为计算机很难擅长做数学。

6. 数学研究何以可能进行？

反过来，我们也可以问，数学研究可能进行，这里面哪一点显得很奇怪呢？我在本书中提到过几个未解决的问题，而数学研究在很大程度上就是在努力解决这些问题以及类似问题。如果读过了第七章，那你会看到，有一种生成问题的好办法是去找一种很难精确分析的数学现象，然后努力对它作一些近似的陈述。第六章结尾处还提出了另一种办法：选一种较难的数学概念，诸如四维流形，然后你通常就会发现，关于这些概念，即便很简单的问题也非常难解答。

如果说数学研究中有什么比较神秘的话，那并不是困难问题的存在——实际上，要创造出奇难无比的问题很容易——而是居然有足够的问题恰好有相当的难度，从而钓住数以千计的数学家。要实现这一点，这些问题必然得有挑战性，但同时也必须让人们看到有可能解决的一线希望。

7. 有没有著名数学问题被业余爱好者解决过？

坦率地讲，没有——这就是对这个问题最简单的回答，也是最不具有误导性的回答。专业数学家能够很快地意识到，他们就著名问题所产生的几乎任何思想，都已经有许多前辈想到过了。一种思想要想成为全新的，就必须具备某种特征能够说明为何前人从来没有考虑过它。可能仅仅是这种想法极具原创性，出人意料，但这种情况十分罕见：总体而言，某种思想的诞生会有充足的理由，而不会是凭空冒出来的。如果你有了这种想法，那凭什么别人就不曾有过呢？一种更加合理的理由是，这个想法和别的某种思想相关，那种思想的知名度并不高，而你已经不畏艰难地去学习并且吸收那种思想。这样至少降低了别人在你之前已经有过同样想法的概率，虽然还是没有降到零。

世界各地的数学系所经常收到一些信件，寄信人声称他们解决了某某著名问题，但实际上无一例外，这些“解答”不仅是错误的，而且错得还很滑稽。有一些解答虽然严格地说并不是犯了错误，看起来却根本就不像是在正确地证明任何东西，简直都不能算是去尝试解答。有些解答至少遵循了数学陈述的常规方式，但其中使用的都是非常初等的论据，如果这些论据是正确的，早在许多世纪以前就会被人发现了。写这些信的人对数学研究的艰难程度没有一点概念，不了解要想做出重大的原创性工作，必须要花上数年的时间来充分发展知识和专长，也不知道数学是一项多么需要集体合作的活动。

上面最后一句话并不是指数学家会在很大的团体中工作——尽管很多研究论文的确都署有两到三个作者的名字。我所指的是，随着数学的发展，人们创造出解答某些问题不可或缺的新技术。其结果就是，每一代数学家都站在前一代数学家的肩膀上，去解答过去一度认为难以触及的问题。如果你试图离开数学界主流去孤立地研究问题，那你就必须自行去创造这些技术，这当然会使你处于一种极其不利的境地。

当然我也不是说业余爱好者就不可能在数学中作出重要的研究。实际上的确有一两个这样的例子。1975年，一位几乎没有接受过任何数学训练的圣地亚哥家庭主妇玛乔丽·赖斯，在读过《科学美国人》杂志中谈到的问题后，发现了三种之前不为人知的用（非正）五边形镶嵌平面的办法。在1952年，一位58岁的德国校长库尔特·黑格纳证明了高斯遗留下来一个多世纪的一个著名猜想。

然而，这些例子与我之前所说的并不矛盾。有一些问题看起来与数学的主干联系并不太紧密，对于这些问题来讲，了解现有的数学技术帮助就不是很大。寻找新的五边形镶嵌方式的问题就属于这样的情况：专业数学家去解决它和一位有天资的爱好者去解决它，两者的水平几乎不相上下。赖斯所取得的成绩就好像是一位天文爱好者发现了一颗新的彗星，解决问题为她带来的名声正是她长久研究所应得的回报。至于黑格纳，尽管他不是一位专业数学家，他肯定也不是在孤立地开展工作。尤其是，他自学了模函数的知识。我无法在此解释模函数是什么——实际上，这部分知识通常被视为比数学系的本科课程还要高深。

有意思的是，黑格纳并没有通过完全常规的方式详细写下他的证明。而且尽管他的论文勉强发表了，在很长的时间里也都被当作是错误的。直到1960年代末，这个问题由艾伦·贝克和哈罗德·斯塔克独立地再

次解决。直至此时，黑格纳的工作才被仔细地重新检视，并被发现是正确的。不幸的是，黑格纳已经在1965年去世，没能亲眼看到自己恢复名誉的那一天。

8. 数学家们为什么会认为某些定理和证明是美丽的？

在本书前面的部分我讨论过这个问题，所以在这里我会说得很简要。用美学的语言来表述数学这一类明显枯燥的事物，这似乎有些奇怪。但正如我在第三章中（在铺地砖问题的结尾处）所说明的，数学论述能够给人愉悦感，这样的愉悦感与更传统的美学愉悦感有很多共同点。

不过，其中一个不同点是——至少在美学观点看来——数学家比艺术家缺少个人特质。我们可能会极为景仰某位发现了美丽证明的数学家，但这项发现背后的人的故事却会最终被淡忘，结果能够愉悦我们的还是数学本身。

索引

（条目后的数字为原文页码）

A

abstract method， the 抽象方法18，20——26， 29——34， 62， 65， 70——71， 82， 84——85， 105， 133

abstraction 抽象16

approximation 近似56——57， 61， 63——64， 66——69， 114——115， 121——122， 124

of logarithms 对数的116——117

of multiplication 乘法的116

of square roots 平方根的116

Archimedes 阿基米德64， 68

area 面积64——69， 82

of a circle 圆的68——69

of curvilinear shapes 曲线形状的 65——67

of polygons 多边形的64——65

associative law， the 结合律23， 32——33， 62

atlas 地图册

concept of 的概念109——111

of a torus 环面的109——110

axioms 公理39——41， 86

for addition and multiplication 加法和乘法23， 27

for area 面积65

for inverses 倒数26——27

for logarithms 对数34

for powers 指数33

of Euclidean geometry 欧氏几何 86——87， 90——91， 93， 97——98， 102

B

beauty in mathematics 数学之美51， 137——138

c

cancellation law， the 消去律27
Cantor， Georg 格奥尔格·康托尔40
Cartesian coordinates 笛卡尔坐标 71——74， 92， 97， 106
chess 国际象棋19——20
commutative law， the 交换律23， 62
complex numbers 复数30——31
computers 计算机11——114， 122， 134
consistency， importance of 自洽性的重要性32， 41， 71
curved space 弯曲空间104——108

D

decimal expansions 十进制展开30， 57——62
Descartes， René勒内·笛卡尔71
dimension 维度see high-dimensional geometry 参见高维几何
dislike of mathematics 厌恶数学 131——132
distance 距离73——75， 98——103
distributive law， the 分配律23

E

Eddington， Sir Arthur 阿瑟·爱丁顿爵士107
Einstein， Albert 阿尔伯特·爱因斯坦 107
Euclid 欧几里得56， 86， 90
Euclidean geometry 欧氏几何see geometry，Euclidean 参见几何，欧氏
Euclid’s Elements欧几里得的《几何原本》86
exponentiation 指数幂117

F

Fermat’s Last Theorem 费马大定理 128
fractional dimension 分数维度82——85
fractions 分数see rational numbers 参见有理数
Frege， Gottlob 戈特洛布·弗雷格40
fundamental theorem of algebra， the 代数基本定理31
fundamental theorem of arithmetic， the 算术基本定理52

G

Galileo 伽利略3
Gauss， Carl Friedrich 卡尔·弗雷德里希·高斯31， 90，107， 137
genius 天才127——128
geometry 几何
Euclidean 欧氏86——91
hyperbolic 双曲98——104，107——108
spherical 球面93——98
Goldbach's conjecture 哥德巴赫猜想121
golden ratio， the 黄金分割比42——45
graphs 图15——16， 19
great circles 大圆93——95， 97

H

Heegner， Kurt 库尔特·黑格纳137
high-dimensional geometry 高维几何 70——80
cubes and spheres in 中的立方体与球面75——79
reason for studying 研究的理由 80——81
Hilbert， David 大卫·希尔伯特ix
Hilbert spaces 希尔伯特空间ix——x
hyperbolic geometry 双曲几何see geometry，hyperbolic 参见几何，双曲

I

induction， principle of 归纳法原理 39， 41
infinity 无穷32， 56——62， 64
instantaneous speed 瞬时速度62——64
irrational numbers 无理数29， 36

J

Jordan curve theorem， the 若尔当曲线定理54

K

Kant， Immanuel 伊曼努尔·康德90
Koch snowflake， the 科赫雪花83——85

L

logarithms 对数34， 116——117， 119， 124—— 125

M

manifolds 流形108——111
mathematical models 数学模型3——5，
16， 28， 31， 71， 80
of brains 大脑的14
of computers 计算机的11——14
of dice 掷骰子的5——6
of gases 气体的8——11
of high-dimensional space 高维空间的71——75
of population growth 人口增长的 6——7
of timetabling 时间表的14——15
Mercator’s projection 墨卡托投影 100——101
music and mathematics 音乐与数学 51， 130——131

N

natural numbers 自然数20——24
negative numbers 负数26——28
number systems 数系21——22， 26， 28——29

P

parallel postulate， the 平行公设87—— 88， 102——104
attempted proofs of 的尝试证明 91——98，104
Poincaré， Henri 亨利·庞加莱98
Poincaréconjecture， the 庞加莱猜想111
prime number 素数118——122
theorem， the 定理119
probability 概率5——6， 8， 10，118——121
proof 证明36， 39——42， 48， 50——51
by contradiction 反证法36
of Pythagoras’theorem 毕达哥拉斯定理的48——49
that the golden ratio is irrational 黄金分割比为无理数的41——45
that the square root of two is irrational 根号二为无理数的 36——39
Pythagoras 毕达哥拉斯29
theorem of 定理48——49， 73——74

Q

qualia 感受性31
Quicksort 快速排序125

R

raising numbers to powers 指数运算 33——34， 133
rational numbers 有理数26——28，36——37
real numbers 实数29
reductio ad absurdum 归谬法see proof by contradiction 参见证明——反证法
reinterpretation 重新解释97——98， 103
Rice， Marjorie 玛乔丽·赖斯136——137
Riemann hypothesis， the 黎曼假设 121——122
Russell， Bertrand 伯特兰·罗素40， 140

S

Saussure， Ferdinand de 费尔迪南·德·索绪尔18， 139

simplifying assumptions 简化假设2—— 3， 6——9

sorting algorithms 排序算法122——125

spherical geometry 球面几何see geometry，spherical 参见几何，球面

square root of minus one， the 负一的平方根see complex numbers 参见复数

square root of two， the 根号二29—— 30， 36——38， 56——62

T

teaching of mathematics 数学教育 132——133
tiling problem 铺地砖问题49——51
trefoil knot， the 三叶结52——53

V

Vinogradov， Ivan 伊万·维诺格拉多夫120——121
visualization 图像化78——80
volume 体积82——83

W

Whitehead， Alfred North 阿尔弗雷德·诺斯·怀特海40，140
Wiles， Andrew 安德鲁·怀尔斯127——128
Wittgenstein， Ludwig 路德维希·维特根斯坦18， 28，139
women in mathematics 数学领域的女性128——130， 140

Z

zero 零24——25
division by 作除数26， 32